FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Amostragem e Caraterização de Coleções de Dados do Twitter

Tiago Manuel da Silva Barreiro de Magalhães

Mestrado Integrado em Engenharia Informática e Computação

Orientador: Sérgio Sobral Nunes (Prof. Auxiliar)

Junho de 2013

Amostragem e Caraterização de Coleções de Dados do Twitter

Tiago Manuel da Silva Barreiro de Magalhães

Mestrado Integrado em Engenharia Informática e Computação

Aprovado em provas públicas pelo júri:

Presidente: Luís Filipe Pinto de Almeida Teixeira (Prof. Auxiliar)
Vogal Externo: Nuno Filipe Fonseca Vasconcelos Escudeiro (Prof. Adjunto)
Orientador: Sérgio Sobral Nunes (Prof. Auxiliar)

18 de Julho de 2013

Resumo

A crescente utilização de redes sociais, de que é exemplo o Twitter, tem vindo a atrair o interesse científico, originando diversos estudos e projetos de investigação acerca dos hábitos e características das suas comunidades. Este fenómeno deve-se essencialmente às suas principais propriedades, como a rapidez e facilidade de publicação de pequenas mensagens de 140 caracteres (*tweets*) apartir de qualquer parte do mundo. Além dos estudos sociais e estatísticos, caracterizadores dessas comunidades, o Twitter serviu igualmente de base para estabelecer paralelismos entre o que aí ocorre e o que poderá refletir do mundo real. No entanto, muitos destes estudos ignoram os possíveis enviesamentos que a obtenção destes dados poderá originar, colocando em risco a validade das conclusões.

O principal objetivo deste trabalho foi investigar sobre as diferenças entre as amostras obtidas através de diversos tipos de extração de dados do Twitter. Os métodos foram implementados para obter diferentes amostras representativas de cada recolha. Optou-se por estudar quatro métodos de recolha de utilizadores diferentes: recolha por IDs de utilizador gerados aleatoriamente; recolha pelos IDs de utilizador presentes nas listas de "seguidores"; recolha pelos IDs de utilizador presentes nas listas de "seguidos"; recolha dos autores dos *tweets* presentes na *sample stream*, a amostra do fluxo de dados cedida gratuitamente pelo Twitter, que corresponde a 1% da totalidade do fluxo público de *tweets*. Para cada amostra, procedeu-se a uma caracterização estatística dos seus dados, relativos aos atributos mais importantes dos seus utilizadores, assim como aos mais estudados. Através da comparação entre os dados das suas amostras, foi possível observar diferenças e semelhanças entre as mesmas. As diferenças entre as amostras são significativas. Por exemplo, através da recolha de utilizadores obtidos pela *sample stream* foi extraída uma amostra que revela uma média de *tweets* publicados pelos seus utilizadores superior às restantes. Por outro lado, através da exploração das listas de "seguidos", foi possível recolher utilizadores com maior número de utilizadores presentes nas listas de "seguidos"e "seguidores". Concluíu-se que a forma como retiramos dados do Twitter influência o tipo de amostra que obtemos para posterior análise. Deste modo, qualquer trabalho que tenha como base dados obtidos da rede social Twitter deve realçar todos os aspetos relativos à recolha de amostras, assim como os possíveis enviesamentos que as mesmas possam originar.

Abstract

The growing usage of social networks, Twitter being one of them, has atracted the scientific interest, giving birth to several studies and investigational projects about the habits and caracteristics of its communities. This phenomenon is mainly due to its major properties, such as the quickness and ease in the publication of small messages 140 caracters long (*tweets*) from anywhere arround the world. Besides the social and statistical studies, which characterize these communities, Twitter also served as a basis for establishing paralels between what happens there and what might reflect in the real world. However, many of these studies ignore the possible bias wich these data may lead, endangering the validity of the conclusions. The main objective of this work was to investigate the diferences between the samples retrieved through various kinds of data extraction of Twitter. The methods were implemented to obtain representative samples from each collection. It was decided to study four diferent user collection methods: collection of user IDs randomly generated; collection of users from the followers lists; collection of users from the friends list; collection of tweet authors on the sample stream, the sample of the data stream freely given by Twitter, that corresponds to 1% of the total public tweet stream data. To each sample, it has been performed a statistical characterization of its data, for the most important fields of its users, as well as the most studied. By comparing the data of the samples, it was possible to observe the diferences and similarities between them. The diferences among samples are significative. For example, the data from the sample stream reveals users with an average of tweets higher than the rest of the samples. On the other hand, by exploring the friends list, it was possible to retrieve users with greater number of users in their friends and followers lists. It has been concluded that the way we collect data from Twitter influence the type of data that we obtain for further analysis. Thus, any work that is based on data retrieve from the Twitter social network must highlight all the aspects relating the data collection, as well as the possible biases that they may lead.

Agradecimentos

Agradeço ao meu orientador, o Prof. Sérgio Nunes, pelo seu tempo, dedicação e paciência no tempo pelo qual me encontrei a elaborar o presente trabalho. Seguidamente, agradeço do fundo do coração à Maria Camps, cujo apoio moral se revelou indispensável para a finalização deste documento, assim como todo o apoio prestado pela minha família. Para finalizar, os meus sinceros agradecimentos a todos que tomaram algum do seu tempo para me esclarecer e debater aspetos do meu trabalho, nomeadamente ao Álvaro Monteiro, Mário Carneiro, Hélder Tavares e a João Gradim.

Tiago Manuel da Silva Barreiro de Magalhães

Conteúdo

1 Introdução **1**
- 1.1 Contexto . . . 1
- 1.2 Objetivos e Motivação . . . 2
- 1.3 Organização do documento . . . 2

2 Amostragem de Dados do Twitter **3**
- 2.1 Introdução . . . 3
  - 2.1.1 Twitter Social Network . . . 3
  - 2.1.2 Utilizadores no Twitter . . . 4
  - 2.1.3 Tweets . . . 6
- 2.2 Extrair dados do Twitter . . . 7
  - 2.2.1 Extração de dados com base nos utilizadores . . . 9
  - 2.2.2 Extração de dados com base em pesquisas de Tweets . . . 11
  - 2.2.3 Extração de dados com base na *Public Timeline* . . . 13
  - 2.2.4 Extração de dados com base no ID . . . 14
  - 2.2.5 Sistema Automático . . . 15
  - 2.2.6 Serviços e Coleções Externas . . . 16
  - 2.2.7 Qualidade da Amostragem . . . 16
- 2.3 Conclusão . . . 18

3 Metodologias para Extração de Dados no Twitter **21**
- 3.1 Considerações Gerais . . . 22
- 3.2 Introdução aos métodos das APIs do Twitter . . . 22
- 3.3 Exploração das ligações entre utilizadores . . . 24
- 3.4 Números de ID aleatórios . . . 26
- 3.5 Dados provenientes da Stream . . . 27
- 3.6 Representação dos dados . . . 27

4 Análise de Dados **29**
- 4.1 Dados gerais das recolhas . . . 29
  - 4.1.1 Comportamento dos métodos . . . 30
- 4.2 Utilizadores . . . 32
  - 4.2.1 Data de Registo de utilizador . . . 32
  - 4.2.2 Tweets . . . 33
  - 4.2.3 Média diária de Tweets . . . 37
  - 4.2.4 Seguidores . . . 38
  - 4.2.5 Seguidos . . . 41

4.2.6 Língua 44
4.2.7 Privacidade 45
4.2.8 Localização 45
4.2.9 URL 46
4.2.10 Fuso horário 47
4.2.11 Geolocalização 48
4.2.12 Verificação 48
4.2.13 Cobertura de dados 49
4.2.14 Intervalos de Confiança 50

5 Conclusões e Trabalho Futuro **51**
5.1 Conclusões 51
5.2 Trabalho Futuro 53

Referências **55**

Lista de Figuras

4.1 Entradas de dados na Base de Dados por hora verificado nos métodos. . . 31
4.2 Evolução dos registos agrupados por ano. 33
4.3 Data dos *tweets* emitidos mais recentemente pelos utilizadores das várias amostras, agrupados por ano. 34
4.4 Total de tweets emitidos por ano pelos utilizadores recolhidos. 35
4.5 Diagrama de caixa e fio relativa aos valores do campo de total dos tweets emitidos por utilizador. 36
4.6 Diagrama de caixa e fio relativa aos valores das médias diárias de *tweets* publicados pelos utilizadores. 38
4.7 Total de seguidores de utilizadores, por definido intervalos. 39
4.8 Diagrama de caixa e fio relativa aos valores do campo de total dos "seguidores" dos utilizadores. 41
4.9 Total de seguidos de utilizadores, definido por intervalos. 42
4.10 Diagrama de caixa e fio relativa aos valores do campo de total dos "seguidos" dos utilizadores. 43
4.11 Valores do campo *(*lang), correspondente à língua definida para o utilizador. 44
4.12 6 fuso horários mais representados nas amostras de utilizadores. 47

LISTA DE FIGURAS

Lista de Tabelas

2.1 Tabela sumária dos campos associados ao perfil de utilizador. 5
2.2 Tabela sumária dos campos de um *tweet*. 7
2.3 Organização das APIs do Twitter. 8
2.4 Tabela sumária de algumas coleções de dados do Twitter disponíveis na Internet. 16
2.5 Resumo dos vários artigos e trabalhos analisados neste capítulo. 20

3.1 Tabela descritiva dos métodos utilizados no presente trabalho. 23
3.2 Tabela das várias respostas aos pedidos HTTP feitos pelos vários métodos de recolha. 23
3.3 Conjunto base de utilizadores. 24
3.4 Campos do utilizador. 28

4.1 Referências simplificadas aos métodos implementados e às respetivas amostras. 29
4.2 Dados gerais das amostras recolhidas. 30
4.3 Tamanho das amostras normalizadas. 30
4.4 Tabela relativa aos erros retornados por cada invocação do método da API. 31
4.5 Tabela relativa às invocações bem sucedidas de cada método da API. . . . 32
4.6 Tabela sumária dos valores totais de *tweet* de cada utilizador. 36
4.7 Tabela sumária dos valores das médias diárias de utilizador das várias amostras. 37
4.8 Tabela sumária dos valores totais de "seguidores" de cada utilizador. . . . 40
4.9 Tabela sumária dos valores totais de "seguidos" de cada utilizador. 43
4.10 Contagem das contas públicas e privadas de utilizador. 45
4.11 Contagem das contas com localização definida no perfil de utilizador. . . 46
4.12 Contagem das contas com *URL* definido no perfil de utilizador. 46
4.13 Contas de utilizador com geolocalização ativa ou inativa. 48
4.14 Contas de utilizador com verificação e sem verificação. 49
4.15 Número de utilizadores presentes em ambas as amostras. 49
4.16 Relações entre os utilizadores das amostras LSeguidos e LSeguidores. . . 50
4.17 Valores dos intervalos de confiança calculados relativos a diferentes campos de cada amostra. 50

Abreviaturas e Símbolos

API	Application Programming Interface
BD	Base de Dados
IC	Intervalo de Confiança
ID	Identificação
DP	Desvio Padrão
Q	Quartil
URL	Uniform Resource Locator
REST	Representational State Transfer
Var	Variância

Capítulo 1

Introdução

O presente trabalho tem como objetivo a análise das técnicas usadas para recolher amostras de dados do Twitter. O Twitter é uma das redes sociais mais populares e utilizadas do mundo [Ale], e também uma das mais estudadas. Todavia, revela-se como escassa a análise detalhada das amostras recolhidas em vários estudos ou artigos científicos, já que normalmente o objetivo primordial destes se centra mais nas conclusões e inferências analíticas do contéudo da informação agregada na amostra. Dada a inexistência de informação oficial sobre o real fluxo de dados do Twitter, o que dificulta a validação destas amostras como representativas da realidade, serão usados métodos estatísticos para, deste modo, conhecer melhor as amostras e compará-las entre si. Espera-se que, no final deste trabalho, seja possível compreender melhor as referidas técnicas de amostragem e os seus resultados.

1.1 Contexto

Vivemos numa Era rica em quantidade de informação. Desde a recente democratização da Internet que se assistiu a uma crescente partilha de dados voluntária por parte dos seus utilizadores. Como consequência, surgiram novas fontes de informação, assim como diferentes maneiras de recolher da *web*. Entre as várias aplicações que foram emergindo, uma destacou-se pela singularidade e simplicidade: o Twitter, destinado à publicação de pequenos textos de 140 caracteres, rapidamente ganhou notoriedade mundial e é, hoje em dia, utilizado pelos mais diversos quadrantes da sociedade, de formas igualmente distintas e variadas. Foi com naturalidade que se assistiu a um crescente número de estudos e investigações que adotaram como material de amostragem os milhares de dados do Twitter produzidos diariamente. É prática comum neste tipo de investigação atribuir como espelho da realidade a informação estatística proveniente dos grandes conjuntos de dados,

dando aso a conclusões bastante polarizadas [Cra13]. As diferenças entre a realidade e o retrato obtido pelas supracitadas análises começam a ser alvo de mais reflexões por parte da comunidade científica, e é precisamente nesse contexto que este estudo se insere.

1.2 Objetivos e Motivação

O objetivo primordial foi procurar a resposta a uma pergunta simples: recolhas de dados diferentes originam amostras diferentes? Uma amostragem forte e sólida é a base que sustenta qualquer conclusão de um estudo. A informação oficial cedida pelo Twitter acerca do seu fluxo de dados é insuficiente para compreender as características dos seus utilizadores, assim como da comunicação entre si, sendo necessário recorrer aos dados obtidos pelas suas APIs. O presente trabalho pretende analisar as propriedades das várias amostragens obtidas, de forma a tornar claro que tipo de amostra é construída quando extraímos dados do Twitter. Assim, através dum maior conhecimento da estrutura das coleções de dados, será mais fácil não só compreender melhor os resultados apresentados em estudos e investigações do Twitter, assim como os possíveis enviesamentos que as metodologias de recolha de dados podem originar.

1.3 Organização do documento

O documento organiza-se em 5 partes distintas: a introdução ao tema é feita no Capítulo 2 - Amostragem de Dados do Twitter; a descrição dos métodos de extração é apresentada no mesmo capítulo; a recolha dos dados é descrita no Capítulo 3 - Metodologias para Extração de Dados no Twitter; a análise desses dados é apresentada no Capítulo 4 - Análise de Dados; por fim, as conclusões e o trabalho futuro são descritas no Capítulo 5 - Conclusões e Trabalho Futuro.

Na descrição dos métodos de extração, foram identificados os métodos de recolha de dados mais utilizados. Seguidamente, na recolha de dados, foram eleitos e descritos os métodos de extração de dados implementados no presente trabalho, tendo como critério a sua popularidade no meio científico. A posterior análise de dados foi efetuada através de *scripts* desenvolvidos em R, a popular linguagem de programação destinada à análise estatística. Finalmente, nas conclusões e trabalho futuro são descritas as conclusões retiradas do trabalho, assim como o trabalho a efetuar futuramente, dentro do presente tema.

Capítulo 2

Amostragem de Dados do Twitter

2.1 Introdução

Serve o presente Capítulo para descrever e contextualizar a forma e estrutura em que o Twitter se encontra organizado, assim como para apresentar os vários estudos efectuados no seu âmbito, com especial destaque para as metodologias de recolha de dados e informação adotados por cada um. Procura-se identificar os pontos fortes, fracos, assim como aferir a existência de alguma relação entre a técnica adotada e o tipo de resultados (ou objecto de estudo) pretendido.

2.1.1 Twitter Social Network

Antes de identificar os diferentes tipos e métodos de recolha, assim como os estudos e projectos associados a cada um deles, revela-se necessário para a sua compreensão uma introdução ao serviço de microblogging/rede social que é o Twitter.

Desde o seu lançamento oficial em Julho de 2006 até aos dias de hoje, o Twitter tornou-se numa das mais importantes redes sociais presentes na *web* [Nat]. Este serviço permite aos seus utilizadores escreverem e partilharem mensagens de estado com um grupo de seguidores, que podem variar bastante na temática, abordando desde tópicos mais rotineiros e pessoais até notícias ou reacções a eventos de importância social e internacional. Estas actualizações de estado são apelidadas de *tweets*, mensagens com o tamanho máximo de 140 caracteres. Para receber os *tweets* de outros utilizadores na sua área pessoal, um utilizador deverá "seguir" outros, num conceito que difere das relações de amizade de outras redes sociais, já que é possível um utilizador "seguir"outro, sem que o inverso se verifique. Todos os utilizadores dispõem de uma página *web* com a sua área pessoal, onde os seus próprios *tweets* e os de todos os utilizadores contidos no agrupamento denominado "seguindo" (ou seja, utilizadores cujo o utilizador tem interesse em

seguir os seus *tweets*) estão dispostos numa única lista ordenada cronologicamente. De acordo com dados publicados pela empresa francesa Semiocast [Sem], até 1 de Julho de 2012, 517 milhões de utilizadores registaram-se no Twitter, e durante o mesmo mês foram publicados um bilião e 58 mil milhões de *tweets*.

Podemos identificar duas grandes fontes de dados do Twitter: os perfis de utilizador e os *tweets* por estes emitidos. Assim sendo, irão ser descritas as propriedades de cada uma destas fontes, assim como alguns estudos associados e respectivas características.

2.1.2 Utilizadores no Twitter

Os utilizadores do Twitter, ao contrário do que maioritariamente acontece nas redes sociais, têm pouca informação associada ao perfil pessoal. Para além de um nome de conta (o *username*, que tem que ser único), um utilizador pode também definir um nome, endereço de correio eletrónico, localização, endereço de uma página *web* pessoal e uma pequena biografia de 140 caracteres (a mesma dimensão de um *tweet*). Dos dados atrás enumerados, apenas o nome de utilizador é obrigatório. No campo de localização, por exemplo, é permitido ao utilizador preencher com qualquer tipo de combinação de caracteres, o que dificulta a identificação geográfica dos utilizadores através do seu perfil. Ao mesmo tempo é permitido ao utilizador recorrer a este atributo para se exprimir, de que são exemplos "Wonderland" (um local imaginário) [CCL10] ou "America i wish, England:(" [TGW12], exibindo desejos ou anseios pessoais, ou até mesmo fazendo referência a múltiplas localidades, como "From Dallas but live in ATL.". Ainda no campo da localização, o utilizador também pode activar a opção de *geotagging*, que revela a localização de onde o *tweet* está a ser emitido, e resulta na anexação dessa informação ao próprio *tweet*. As propriedades deste atributo serão explicadas com maior detalhe na secção 2.1.3. Finalmente, o utilizador pode definir se deseja ter um perfil privado, implicando que a partilha de *tweets* irá ocorrer exclusivamente com os seus seguidores, inibindo outros utilizadores da comunidade de visualizarem os mesmos. Accionando esta opção, todos os *tweets* serão excluídos de pesquisas que estejam relacionadas com a *Public Timeline*, a coleção de todos os *tweets* públicos construída em tempo real.

Existem várias formas de interação entre os utilizadores do Twitter. Além da criação de atualizações de estado (*tweets*), mencionada anteriormente, que permite a visualização dos *tweets* criados por cada utilizador, outros hábitos começaram a ser desenvolvidos na sequência do uso da rede social. Os mais populares acabaram por definir os métodos de comunicação do Twitter, tornando-se em funcionalidades da aplicação de forma oficial.

Um dos hábitos adotados pelo Twitter como mecanismo de comunicação no seu seio foi a utilização do caracter "@" para indicar "addressivity", ou seja, para endereçar os *tweets* a um ou mais utilizadores, permitindo discussões abertas e facilitando a sua actualização. Herring et al. [HH09] exploraram os aspectos e características da colaboração no

Twitter, concluindo que, apesar do Twitter não estar desenhado de raiz para esse fim, as conversações no Twitter vieram a ser mais recorrentes ao longo do tempo.

Outra forma comum de interação entre utilizadores no Twitter prende-se com o uso de *retweets*. Os *retweets* são reproduções de *tweets* de outros utilizadores. Como observado por Suh et al. [SHPC10], analisar esta funcionalidade pode ser algo difícil, já que os *retweets* não seguem um padrão comum na sua produção, variando se emitidos a partir das diferentes aplicações cliente ou do próprio *site*. Entre utilizadores que partilham uma relação recíproca, ou seja, que se "seguem", é permitida a troca de mensagens directas e privadas, bastante semelhantes a uma SMS. Esta funcionalidade do Twitter é muito utilizada por utilizadores que acedem ao Twitter através dos seus dispositivos móveis, já que substitui o envio de SMS, um serviço geralmente pago.

Sendo o Twitter uma rede social, as suas relações podem ser definidas como um grafo social. Como os utilizadores estão directamente ligados pela relação de "segue", é possível definir estas relações num grafo, tal como G(V,E), onde V é o conjunto de utilizadores e E o conjunto de utilizadores que seguem esses utilizadores [JSFT07]. No entanto, trata-se de um grafo com características muito próprias. Como anteriormente mencionado, um utilizador tem uma lista de diferentes "seguidores", assim como outra dos utilizadores que vai "seguindo". Na análise da relação entre dois utilizadores, é crucial identificar a existência, ou não, de reciprocidade entre os mesmos. Este fator define os 3 tipos de relação existentes no Twitter (assumindo *a* e *b* como dois utilizadores distintos): *a* segue *b*; *b* segue *a*; *a* e *b* seguem-se mutuamente. A análise desta rede tem sido alvo de diferentes estudos, sendo também usada em algumas técnicas de recolha de dados do Twitter, que serão alvo de uma maior análise no presente documento.

Campo	Descrição
Nome de conta	Nome identificador da conta de utilizador, único mas passivo de modificação pelo utilizador
Nome	Nome do utilizador
Descrição	Descrição da conta do utilizador
Contagem de favoritos	Número de *tweets* que o utilizador definiu como favoritos
Língua	Língua do perfil do utilizador
Localizaçao	Campo destinado à localização, a preencher pelo utilizador
Privacidade	Informação sobre a privacidade dos dados da conta e *tweets*
Contagem de *tweets*	Número de *tweets* publicados
Fuso horário	Fuso horário definido pelo utilizador
URL	Campo reservado para uma hiperligação

Tabela 2.1: Tabela sumária dos campos associados ao perfil de utilizador.

2.1.3 Tweets

Os *tweets* constituem a fonte mais dinâmica e diversa no tipo e qualidade de informação que se encontra disponível para recolha. Um *tweet* caracteriza-se como tendo um contéudo textual limitado a 140 caracteres, como mencionado no ponto 2.1.1, mas é importante analisar as restantes propriedades que o definem. Existem símbolos que são encontrados recorrentemente, pois possuem significados e utilidades específicas. Como mencionado no ponto 2.1.2, o símbolo "@" denomina "addressivity", mas não é o único que foi aproveitado pela comunidade e pela aplicação como mecanismo de referência. Também o símbolo "#" determina o tema ou rótulo (p.ex. "#rtp" para identificar um *tweet* como relativo ao canal de televisão RTP), sendo uma medida de conversação já largamente adoptada pela comunidade e integrada quer nas aplicações clientes de Twitter, quer no próprio site do Twitter.

Um dos conceitos introduzidos pelo Twitter centra-se na aglomeração de informação contextual de um *tweet*, sob a denominação de Entidades [Twil]. As entidades podem ser relativas a informação sobre contéudos multimédia, a *URLs* presentes no *tweet*,a menções a utilizadores e também a *hashtags*, e visam simplificar o trabalho de processamento e análise do texto por parte do investigador.

Com a introdução das entidades, o Twitter possibilita ao programador retirar mais pedaços de informação pré-tratada, referente ao seu contéudo. No entanto, normalmente os estudos recolhidos preferem desenvolver os seus próprios métodos de processamento de texto, controlando todos os procedimentos de tratamento de informação. O fato de ser um tipo de informação relativamente recente no Twitter também explica a sua pouca utilização.

Para além do texto do *tweet*, existem outros campos que podem ser utilizados para a extração de conhecimento do Twitter, já que um *tweet* contém também informação relativa ao seu autor, à data da sua criação, coordenadas geográficas, se se trata duma resposta a outro *tweet*, entre outras propriedades menos relevantes. As coordenadas geográficas apresentam-se como um *geocode*, que é um par de números reais, correspondentes à latitude e longitude. Na tabela 2.2 podemos ver o sumário destes atributos.

Campo	Descrição
autor	autor do *tweet*
texto	corpo do *tweet*
data de criação	data de publicação do *tweet*
coordenadas geográficas	*geocode* referente ao *tweet*
resposta a *tweet*	se é uma resposta a um *tweet*, representado pelo *id*
destinatário	no caso de se verificar que é uma resposta, este campo representa o autor do *tweet* anterior
retweet	se é um *retweet*
entidades	entidades presentes no *tweet*
contagem de favorito	indicador aproximado do número de utilizadores que selecionou o *tweet* como favorito
língua	língua detetada automaticamente no *tweet*
place	associação do *tweet* a um lugar
contagem de *retweet*	número de vezes que o *tweet* foi publicado por outros utilizadores, através de um *retweet*
fonte	fonte do *tweet*

Tabela 2.2: Tabela sumária dos campos de um *tweet*.

2.2 Extrair dados do Twitter

Nesta secção será descrito como os investigadores, programadores e demais interessados utilizam os seus recursos para extrair dados e informação dos contéudos criados no Twitter.

Ao longo dos parágrafos da secção 2.1, verificámos o contributo dado pelo Twitter à forma como interagimos na *web*, não só analisando o tipo de informação que se encontra disponível nos seus utilizadores e *tweets*, mas também na forma como as pessoas comunicam entre si. Uma aplicação que, ao longo dos anos, modificou os hábitos de partilha de informação, e moldou-se ao uso dos seus utilizadores e às tendências tecnológicas que as acompanharam, tal como o crescente uso dos *smartphones* ou *tablets*. Devido à sua natureza de partilha imediata, numa possível e teórica aproximação ao que as pessoas pensam e como se comportam, a comunidade ciêntifica obviamente adotou, entre outros serviços, o Twitter como base de estudo, sob os mais variados prismas. As características relacionadas com as redes sociais que fazem parte da forma como o Twitter se encontra organizado, onde existe uma forte interação entre utilizadores, permitiu originar diversos estudos sobre a forma como as pessoas comunicam entre si. A mensagem de texto incluida num *tweet* permite a diversos investigadores aproveitarem o Twitter para efectuar inúmeros estudos, como por exemplo análises de sentimento, relacionados com marcas publicitárias ou até catástrofes e entretenimento. Outro aspecto bastante explorado no Twitter é a sua capacidade de revelar a posição geográfica dos seus utilizadores, suscitando a atenção dos investigadores para esse tipo de trabalho, desde o aproveitamento dos

Twitter APIs		
REST API	REST API	Métodos globais da *REST API*
	Search API	Métodos de pesquisa de *tweets*
Stream API	Public Streams	*Tweets* públicos
	Site Streams	*Tweets* de vários utilizadores diferentes
	User Streams	*Tweets* do utilizador autenticado

Tabela 2.3: Organização das APIs do Twitter.

atributos de utilizador e *tweets* até ao próprio discurso escrito de um *tweet*. Trata-se de uma aplicação passível de ser utilizada em diferentes áreas de investigação, dada a heterogeneidade relativa não só ao seu uso, como à diversidade do tipo de informação que podemos recolher.

Uma das ferramentas chave para a extração de dados no Twitter é a própria API [Twio]. As desvantagens das metodologias dependentes da utilização da API do Twitter, nomeadamente da Search API [Twii] e da Stream API [Twij], são os limites impostos pela própria empresa nos pedidos aos servidores, o que implica uma preocupação adicional por parte dos investigadores para evitar a proibição temporária do uso das funcionalidades da API para os seus projectos. Para evitar os limites impostos, é frequente o recurso a várias máquinas ou contas de Twitter, para desse modo permitir mais pedidos por hora, reduzindo o tempo de espera. Na tabela 2.3 podemos observar como é que as APIs se encontram organizadas, assim como as suas principais funções.

De acordo com o Twitter, os limites por si impostos afectam as suas APIs de maneiras distintas. Estes variam de acordo com o acesso aos diferentes métodos, explicados com algum detalhe na documentação dirigida aos programadores de aplicações relacionadas com o Twitter [Twin].

Optou-se por dividir os métodos e técnicas com base na fonte de informação utilizada pelos diversos estudos recolhidos. Assim, definiram-se 6 grupos distintos:

- extração de dados com base nos utilizadores;
- extração de dados com base em pesquisas de *Tweets*;
- extração de dados com base na *Public Timeline*;
- extração de dados com base no ID;
- sistema automático;
- serviços e coleções externas.

Primeiramente, irão ser descritas as técnicas associadas à recolha com base na informação associada a uma conta de perfil no Twitter, assim como a forma como os utilizadores se encontram relacionados. De seguida, serão explicadas as metodologias que usam

a *Public Timeline* como matéria prima para a construção das coleções de dados. Outra abordagem consiste em reunir pesquisas de *tweets*, quer no seu texto, quer nos atributos subjacentes. Também o recurso à pesquisa pelos campos de identificação único, quer do utilizador, quer do *tweet*, revelam-se como alternativas para a extração de informação. Para finalizar, existem utilizadores automatizados que reúnem informação da Twittosfera dos seus seguidores, assim como estudos e projectos que, ao invés de terem procedido à extração dos dados através da implementação de métodos da sua autoria, recorreram a recolhas e serviços externos.

2.2.1 Extração de dados com base nos utilizadores

Como referido no ponto 2.1.2, a informação inserida pelo utilizador no seu perfil é escassa quando comparada a outras redes sociais. Contudo, é uma fonte de informação já largamente adotada em diversos estudos sobre o Twitter. As próprias relações entre os seus utilizadores, brevemente apresentadas em parágrafos anteriores, servem não só como métodos de recolha mas como objectos de estudo no universo do Twitter.

Com base num utilizador, podemos aceder aos dados pessoais do seu perfil, às suas listas de ligações com outros utilizadores e à sua *timeline*, que disponibiliza as actualizações de estado dos utilizadores presentes na sua lista de "seguidos".

Um dos métodos baseados na informação existente nos utilizadores assenta na recolha da informação nas ligações de "amizade" que derivam de um ou mais utilizadores iniciais, denominados normalmente por raíz ou base, que sirvam de forma consistente o objectivo da recolha, através de métodos da API. No estudo efectuado por Kwak et al., dedicado à compreensão das características topológicas do Twitter e do seu peso enquanto plataforma de partilha massiva de informação [KLPM10], este foi um dos métodos aplicados. Através do perfil da figura pública Perez Hilton, que possuía mais de um milhão de "seguidores" aquando da data da recolha, foi implementada uma pesquisa em largura, tanto na lista de "seguidores" como na lista de "seguindo". Infelizmente não é especificado quantos utilizadores foram recolhidos só com base na recolha iterativa dos utilizadores, mas o número de utilizadores final cifrou-se nos 41,7 milhões, quando completado com uma recolha de perfis que referiram nas suas actualizações de estado "assuntos do momento" (tópicos mais populares durante um período de tempo), que ocorreu durante um mês.

O ponto de partida para estas recolhas, da qual Perez Hilton fez parte no exemplo acima referido, pode não ser pré-determinado pelos investigadores. A determinação do ponto de partida pode ser definido através de uma pesquisa auxiliar. Para assegurar que os utilizadores que formam a base são utilizadores activos, uma opção popular consiste em recolher perfis directamente da *Public Timeline*, para em seguida, à semelhança da técnica acima descrita, percorrer os utilizadores relacionados com estes e, num passo seguinte na

iteração, também percorrer as listas dos mesmos, como é descrito por Krishnamurthy, Gill e Arlitt [KGA08].

Para a filtragem com base nos dados disponíveis pelos perfis de utilizador, o que assume mais destaque é o campo de localização, utilizado para a filtragem geográfica dos utilizadores. Este método requer normalmente a existência de um dicionário de palavras chave que desejamos filtrar, que associado a técnicas normalmente ligadas a expressões regulares, identifica quais os indivíduos adequados ao estudo. Também a hora de publicação de um *tweet* é usada para ajudar a filtrar a sua geografia. No estudo "Geographic Dissection of the Twitter Network", conduzido por Kulshrestha et al., relativo às ligações criadas pelas pessoas do mesmo sítio e como estas foram afectadas pela existência de redes sociais como o Twitter, estes dois campos para atribuição de nacionalidade dos utilizadores foram aproveitados [KKNG12].

De ressalvar que, em certas situações, quando a base de utilizadores já está identificada, e quando o estudo exige uma monitorização dos *tweets* por parte de um ou mais utilizadores (p.ex. para estudar a frequência e o tipo de uso de *tweets* dum determinado grupo), é possível recorrer às *User Streams* [Twip], que são os agrupamentos das actualizações do próprio utilizador juntamente com as actualizações dos utilizadores presentes na lista de "seguindo" de um dado utilizador. Esse método foi aplicado no *TwitterEcho* [BOM+12], onde são guardados os *tweets* de utilizadores portugueses com base neste tipo de monitorização.

Mais objectivamente, o método descrito na presente secção baseia-se, fundamentalmente, na combinação de dois métodos da API: um, para aceder às listas de ids de utilizadores relacionados com um determinado utilizador, e outro para, através das referidas listas, descarregar a informação de utilizador correspondente a cada id. Analisando o pseudo-código descrito no Algoritmo 1, facilmente percebemos como é efetuada a pesquisa por utilizadores.

```
Input: list seed is the list of seed names or ids
uncrawledUsers = seed;
while uncrawledUsers not empty do
    foreach user in uncrawledUsers do
        followersIds = getFollowers(user);
        followers = lookUp(followersIds);
        storeInDb(followers);
        setUserCrawled(user);
    end
    uncrawledUsers = readFromDb();
end
```

Algoritmo 1: Pseudo-código de uma possível abordagem à recolha de utilizadores.

Numa primeira fase, é necessário establecer o conjunto de utilizadores base para pesquisar, adequando-o ao tema ou objectivo do estudo. Este conjunto será o primeiro a ser analisado, através da combinação de métodos da API que retornam as listas de amigos, *friends/ids* [Twic], ou seguidores, *followers/ids* [Twib], e o método *users/lookup* [Twig] para extrair toda a informação relativa aos utilizadores. Para verificar um utilizador uma única vez, é necessário modificar o seu indicador de estado (se visitado, se por visitar), neste caso representado pela função *setUserCrawled(user)*, que efectua essa mudança após os registos relativos aos seguidores ou amigos serem devidamente guardados.

2.2.2 Extração de dados com base em pesquisas de Tweets

Neste ponto será discutida a recolha de dados no Twitter com base em pesquisas por *tweets* cujos parâmetros e propriedades foram definidos pelos investigadores. Através de um ou mais campos, várias pesquisas são efetuadas, realizando de seguinda a estruturação dos resultados dessa mesma pesquisa e utilizando os recursos associados à *Search API* ou *Stream API*, em conjunto com os métodos da *REST API*.

No estudo "A qualitative examination of topical tweet and retweet practices", de Nagarajan et al., realizado sobre a prática e uso de *tweets* e *retweets* acerca de um determinado tópico [NPS10], foram usados métodos da Search API, tendo como ponto da partida uma base composta por várias palavras-chave relacionadas com três eventos distintos (eleições no Irão, plano de reforma de saúde proposta por Obama e a International Semantic Web Conference). Estas palavras-chave, escolhidas pelos próprios investigadores, foram determinadas não só para identificar quais os *tweets* relacionados com o evento sob análise, mas também para alargar e melhorar a base, fornecendo consequentemente uma pesquisa mais ampla e abrangente ao estudo. Como é evidente, o objectivo deste estudo visa apenas *tweets* que se relacionam com os referidos eventos, delimitando a amostra pretendida para um subconjunto muito reduzido do Twitter.

Para construir uma recolha com base no contéudo da mensagem do próprio *tweet* é necessário, numa primeira fase, escolher os termos ou palavras chave que irão formar uma base de palavras, que serão utilizados como argumentos na invocação dos métodos de pesquisa. São efectuadas quantas pesquisas o investigador quiser definir, sempre com especial atenção para não ultrapassar os limites definidos na API. Após o retorno das listas de pesquisa, é necessário gravar de forma persistente os resultados.

No Algoritmo 2 está descrito em pseudo-código os passos mencionados anteriormente.

```
Input: integer n is the integer that defines the number of cicles
Input: integer seeds is the list containing the keywords to be used as arguments of
       the search
for i ← 1 to n do
    search = getTweets(seed);
    storeSeeds(search);
end
```

Algoritmo 2: Pseudo-código de uma possível abordagem à recolha com base na pesquisa de *tweets* por texto.

Também para a recolha de utilizadores que partilham os mesmos dados geográficos se recorre a este método, já que existe alguma informação tanto em *tweets* como nos perfis de utilizador quanto à sua localização geográfica. Desde Agosto de 2010 que o Twitter permite aos seus utilizadores associarem nos seus *tweets* a sua localização geográfica, através do geocode (latitude e longitude). Através da *Search API* é possivel filtrar *tweets* que tenham sido criados numa determinada área geográfica, apesar de que, segundo Hecht et al. [HHSC11], em 62 milhões de *tweets* recolhidos, apenas 0,77% continham esses dados. Ainda no tema da utilização de *tweets* para determinar a sua geografia, um dos métodos usados no estudo "The livehoods project: Utilizing social media to understand the dynamics of a city" de Cranshaw et al. [CSHS12] foi a procura por *tweets* relacionados com informação partilhada do Foursquare, a popular rede social com base na localização para *smartphones*, já que muitos utilizadores enviam dados referentes a um *check-in* para o Twitter. Dada a natureza da aplicação, estes *tweets* contêm os dados necessários para a extracção da localização. É contudo interessante verificar que também outras aplicações que enviam notificações para o próprio Twitter podem servir para a extração de dados específicos. Os *tweets* desta natureza simplificam o processo de analisar o contéudo da mensagem, já que estes fornecem de forma organizada e padronizada dados relativos a hábitos diferentes (localização, gosto músical, actividades físicas, entre outros). Para esta técnica, já que se procede à análise do texto de um *tweet*, a implementação segue a lógica descrita pelo Algoritmo 2, em que a base de palavras serão as cadeias de caracteres que a aplicação externa adiciona ao *tweet* (no caso do *foursquare* é adicionado uma hiperligação que começa por "http://4sq.com").

No entanto, o serviço Twitter aconselha aos seus utilizadores a evitarem a *Search API* para a recolha exaustiva de dados utilizando os mesmos parâmetros de pesquisa, sugerindo a utilização da *Stream API* para esse tipo de recolhas [Twid].

Esta API requer uma ligação contínua a um *endpoint*, que irá receber uma amostra da

corrente de *tweets* que estão a ser emitidos, praticamente em tempo real. Já a aplicação dos métodos da *Search API* possibilita pesquisar e obter os *tweets* publicados no passado, num alcance que pode ir até 6 ou 9 dias [Twiq].

Num estudo realizado por diversos investigadores com o objectivo de analisar as dinâmicas e a estrutura de uma conversação on-line [BHRG+11] foi utilizada a *Stream API* para recolher os *tweets* relativos ao 15 de Maio (uma manifestação civil ocorrida em Espanha) entre 25/04/2011 e 26/05/2011, obtendo 189 mil *tweets*. Foram aplicados filtros com base em *hashtags* de forma semelhante aos estudos mencionados anteriormente.

Portanto, os dois métodos explicados neste ponto diferem na sua implementação, mas servem o mesmo propósito: obter amostras de *tweets* através da filtragem dos mesmos com base em *hashtags*, palavras, ou outros atributos (localização, língua, entre outros).

2.2.3 Extração de dados com base na *Public Timeline*

Outro dos métodos amplamente utilizados para a extracção de dados do Twitter baseia-se na recolha dos *tweets* diretamente da *Public Timeline*, através de métodos definidos presentemente na *Stream API* [Twij] e, embora agora descontínuado, no método *statuses/public* [Twie] da API do Twitter. De salientar que também o *Twitter Firehose* [Twim], agora apenas acessível a parceiros da empresa, era usado para recolher *tweets* da *Public Timeline* do Twitter. Utilizando a *Stream API* é possível "escutar" alguns dos *tweets* que estão a ser criados no momento, quer do ponto de vista de um utilizador, quer do ponto de vista global. No primeiro caso, "escutamos"apenas os *tweets* relativos aos seus "seguidores" e as menções direcionadas ao utilizador; no segundo caso, recebemos os *tweets* recentes e de forma aleatória, já que o Twitter nunca explicita como é feita a referida seleção. Normalmente, este tipo de recolha consiste em aplicar os métodos supracitados ao longo de semanas, meses ou mesmo anos, guardando de forma persistente os dados obtidos, e cuidadosamente equilibrar os pedidos ao servidor, já que estes se encontram limitados pelo Twitter. Como se trata duma amostra da *Public Timeline* do Twitter, uma das grandes vantagens para um estudo do uso corrente do Twitter é que apenas utilizadores activos serão filtrados, já que utilizadores que não actualizam o seu estado não aparecerão, obviamente, na recolha. Por norma, a maioria da informação recolhida revela-se desnecessária, pois tratam-se de grandes conjuntos de *tweets* em bruto e sem critério. Assim, qualquer estudo que tenha como base a *Public Timeline* não-filtrada do Twitter requer uma fase posterior de tratamento e separação dos dados mais refinada. No estudo sobre hábitos diurnos no Twitter "On the Study of Diurnal Urban Routines on Twitter" [NZBL12], de Naaman et al., foram recolhidos através do *Firehose* os *tweets* entre Maio de 2010 até Maio 2011, identificando posteriormente as localizações dos autores dos *tweets* através do campo de localização presente no perfil. Num outro estudo sobre a interação dos vários tipos de utilizadores no Twitter, "Who says what to whom on twitter" de Wu et

al. [WHMW11], foi usado um dataset recolhido do *Firehose*, durante 223 dias, num total de 5 mil milhões de *tweets*. Neste estudo, apenas 20% da colecção de dados foi aproveitada, em oposição aos restantes 80%, que se revelaram excendentários.

De salientar que, dada a inexistência de outras formas de extrair dados de *tweets*, e possívelmente devido à novidade da aplicação, os investigadores simplesmente denominavam a fonte como "Public Timeline", omitindo o método utilizado para essa extração. No entanto, é mais do que provável que se estejam a referir ao método descontínuado anteriormente mencionado, já que os métodos comportam-se de forma igual, como descrito na documentação do Twitter, obtendo 20 resultados de cada vez.

Num dos primeiros estudos debruçado no Twitter e nas suas propriedades enquanto meio de comunicação, conduzido por Java et al. em 2007 [JSFT07], o método *statuses/public* foi adotado. A pesquisa ocorreu durante um mês, resultando na amostra de 1 348 543 *tweets* pertencentes a 76 177 utilizadores diferentes, recolhendo 20 *tweets* em cada 30 segundos. Para a recolha de perfis aleatórios de utilizadores do Twitter com o objectivo de determinar a forma como a geografia, fronteiras e língua influenciam a interação das pessoas no twitter, Takhteyev et al. em "Geography of Twitter Networks"[TGW12] recorreram à *Public Timeline*, usando o mesmo método, agregando os 20 resultados retornados por cada pedido ao servidor, em intervalos de 25 segundos, durante uma semana em agosto de 2009. Foram recolhidos 481 248 *tweets*, que resultaram numa triagem de 3 360 utilizadores diferentes.

2.2.4 Extração de dados com base no ID

Outro método, menos frequente, prende-se com a utilização do campo de *id* de um *tweet* ou utilizador. Num estudo com o objetivo de clarificar a enumeração dos *ids* dos utilizadores do Twitter [Cli09], em vez de se recorrer a dados públicos do Twitter, foram utilizados métodos da API para pesquisar com base nos ids de utilizador, ou seja, conta a conta, ultrapassando assim as limitações inerentes à pesquisa em dados públicos do Twitter, onde apenas contas públicas são apresentadas. Para tal, criou-se uma nova conta de Twitter e, ciclicamente, gerou-se um grande número de id aleatórios, entre 0 e o número da conta previamente criada (18 496 098), onde se conseguiu filtrar 4 414 contas de utilizador que não retornaram erros de servidor ou de cliente.

No Algoritmo 3 podemos ver um pseudo-código ilustrativo da aplicação desta técnica.

```
Input: integer n that defines the number of cicles
newID = getNewAccountID();
for i ← 0 to n do
    id = generateRandomNumber(newID);
    user = lookUpUser(id);
    if valid(user) == True then
        storeInDb(user);
    end
end
```

Algoritmo 3: Pseudo-código do Algoritmo de recolha de utilizadores pelo campo de ID, de forma aleatória.

Primeiro, uma nova conta é criada, guardando esse valor como o máximo limite numérico que um *id* pode assumir. Depois, ciclicamente, durante *n* iterações, é gerado um novo ID a cada uma, e efectuada uma verificação através do método da API *users/show* [Twih]. Caso esta chamada não retorne erro, é guardada de forma persistente. Caso retorne erro, procede-se à próxima iteração.

2.2.5 Sistema Automático

Outra estratégia para recolher e tratar de dados, e que é aplicada com alguma frequência no universo do Twitter, é a implementação de um utilizador fictício. Apesar de se tratar de uma conta de perfil semelhante às usadas por todos os outros utilizadores comuns, é normalmente mantida por *scripts*, filtrando diretamente a informação dos seus "seguidores". Assim que os utilizadores começam a seguir estas contas *robot*, as mesmas acedem não só aos seus dados de perfil mas também a todas as atualizações de estado, inclusive de utilizadores com contas privadas, já que o acesso a esses dados está dependente da existência da relação "seguir". Os objectivos deste tipo de contas podem ser vários, desde a partilha de notícias relacionadas com um determinado tópico até à análise com base na nacionalidade desses utilizadores. Um popular exemplo do uso de *robots* é o Twitter Portugal [twia], onde são disponibilizados a sua própria *timeline* (supostamente composta apenas por utilizadores portugueses), estatísticas, *tops* de utilizadores, entre outros dados. Este *robot* actualiza o seu estado com periocidade de uma hora, informando qual a *hashtag* mais popular entre os seus utilizadores durante esse tempo. A referida conta tem, até à data presente de 9 de Outubro de 2012, 44 804 utilizadores "seguindo" e 71 884 "seguidores".

2.2.6 Serviços e Coleções Externas

Uma solução diferente consiste em aproveitar recolhas de dados de outros estudos prévios, ou de serviços externos de agregação de conteúdo do Twitter. Destes são exemplos o *Spinn3er* [spi], usado por exemplo num estudo sobre sentimento presente no Twitter [TBP11], ou o GNIP [GNIa], que se tornou na primeira fonte licenciada para a revenda de dados do Twitter desde Novembro de 2010, e que fornece, entre outros serviços, duas feeds de *tweets* diferentes, a Spritzer [GNIb] (que inclui entre 1% a 2% do total de *tweets*) e Decahose [Twik] (10%). Apesar deste ser o método mais fácil e imediato, já que não exige esforço nenhum na sua recolha, permitindo ao investigador mais tempo para a análise concreta de dados, apresenta três potenciais problemas distintos. O primeiro diz respeito ao custo, já que geralmente são serviços privados, e, por isso mesmo, pagos. Quando não o são, são relativamente exclusivos a certas entidades e instituições, nomeadamente académicas; o facto de se tratarem de coleções de dados antigas, com pelo menos meses ou anos, constitui igualmente um problema, já que pode deturpar as conclusões retiradas pelo seu autor; por último, o fato de serem normalmente coleções genéricas e por vezes desajustadas ao objectivo do estudo alvo (que utilizando por exemplo a API, podem ser filtradas numa fase inicial, limitando e adequando a colecção de dados inicial) aumenta o tempo e esforço de análise posterior. Na tabela 2.4 podemos consultar algumas das fontes de informação relativas ao Twitter.

Nome	Endereço	Descrição
TREC	http://trec.nist.gov/data/tweets/	Aproximadamente 16 milhões *Tweets*, recolhidos entre 23 de Janeiro e 8 de Fevereiro de 2011
Infochimps	http://www.infochimps.com/datasets/	Várias amostras de *tweets* e utilizadores diferentes

Tabela 2.4: Tabela sumária de algumas coleções de dados do Twitter disponíveis na Internet.

2.2.7 Qualidade da Amostragem

Nesta secção será descrito um pouco o trabalho já efetuado sobre a amostragem no Twitter quanto à qualidade dos dados recolhidos pelos métodos do Twitter, com especial detalhe no viés, ou tendenciosidade, que estas possam apresentar. Nos trabalhos sobre os quais se esteve a dissertar anteriormente, verificaram e identificaram-se as várias metodologias adotadas para a extração de dados. No entanto, começa a emergir outro tipo de análises, no qual o presente estudo se insere, que pretende verificar se as propriedades das amostragens obtidas com o recurso às metodologias utilizadas pela comunidade científica diferem entre si, e se são uma representação do uso real do Twitter. Como tal, destacam-se dois estudos em particular: "Is the Sample Good Enough? Comparing Data

from Twitter's Streaming API with Twitter's Firehose" [MPLC13], e "Assessing the Bias in Communication Networks Sampled from Twitter" [GBWR+12].

No primeiro trabalho, é efetuada uma comparação entre os dados recolhidos através da amostra cedida gratuitamente pelo Twitter, através da *Stream API*, e os dados obtidos através do *Firehose*, para determinar se os primeiros representam, de uma forma aceitável, o fluxo real do Twitter, que em princípio o acesso pago do *Firehose* garante aos seus programadores ou investigadores. A investigação tomou como campos de análise as *hastags* mais populares presentes nas amostras, a análise de tópico dos *tweets* recolhidos, uma análise à construção da rede feita por utilizadores, ligados através da prática de *retweets* e ainda uma análise à geografia deduzida através de *tweets*.

Para a comparação das *hashtags* mais frequentes nas amostras, foi utilizado pelos investigadores o método estatístico τ de Kendall, entre listas ordenadas de *hashtags* construídas com base nas amostras recolhidas da *Stream API* e do *Firehose*. Concluí-se que, para um maior número de ocorrências, os resultados revelam maior concordância de resultados apresentados pela amostra de *sample*, mas para valores baixos de ocorrências, os dados da *sample* revelaram-se discordantes da amostra do *Firehose*. Para verificar se esse resultado era consequência dos processos de filtragem aplicados pelo Twitter no método de *sample*, foi ainda construída uma amostra de igual tamanho, constituida por *tweets* escolhidos aleatóriamente da amostra do *Firehose*. Curiosamente, para valores baixos de ocorrências, a amostra constituída por dados aleatórios da amostra de *Firehose* revelou-se mais eficiente do que a amostra da *Stream API*, sugerindo que os filtros do *sample* podem originar falácias quanto à analise das *hashtags* mais populares da sua amostra.

Para análise de tópicos dos *tweets*, a equipa de investigação recorreu ao modelo Latent Dirichlet Allocation (LDA), já que se trata de um modelo largamente usado em estudos de Twitter quanto a este tipo de análise. Mais uma vez, a comparação incidiu entre os dados obtidos pelo *sample* da Stream API e subamostras aleátorias pertencentes à amostra recolhida através do *Firehose*. Os investigadores concluíram que, com maior cobertura dos dados, a divergência entre os tópicos de ambas as amostras diminui, constatando também o inverso. Também foram testadas amostras aleatórias pertencentes à amostra total do *Firehose*, que se mostraram eficientes quando usados maiores níves de cobertura de dados. Comparando ambas as amostras (*Stream API* e aleatória), com um limiar de três desvios padrões, concluiram que se obtêm maior concordância entre os tópicos usando as amostras aleatórias.

A equipa de investigadores decidiu analisar as redes de utilizadores criadas através dos *retweets* recolhidos. Analisaram tanto as medidas ao nível do nó da rede, como as medidas quanto ao nível da rede. Concluíram que, através de apenas um dia de extração, é possível identificar, em média 50-60% dos 100 utilizadores-chave. Quanto às medidas ao nível da rede, foi revelado pelo estudo uma correlação entre os índices de centralizaçãoe e a quantidade de dados provinientes da *Streaming API*.

Por fim, analisaram as propriedades dos *tweets* que apresentaram *geotags* de ambas as fontes de dados. Concluiram, comparando ambas as amostras, que a *Stream API* recolhe praticamente todos os *tweets* desta natureza, consequentemente validando este método como uma base forte para os estudos que utilizam os *tweets* com *geolocation*. Concluíram, no final do estudo, que os resultados da *Stream API* dependem fortemente da cobertura de dados e do tipo de analise que o investigador pretende efetuar.

No segundo estudo, pretende-se apurar as diferenças entre amostras obtidas com a *Stream API* e com a *Search API*, assim como as redes de comunicação construídas a partir das mesmas. Para tal, decidiram filtar os *tweets* relativos às manifestações políticas decorridas em Maio, em Espanha, mais precisamente nos dias 12 e 15. A recolha decorreu durante todo o mês. Foram construídas duas redes de comunicação diferentes: uma baseada nas menções a utilizadores, outra nos seus *retweets*.

Numa primeira abordagem, os investigadores descobriram que, embora maior parte dos dados esteja presente em ambas as amostras, existem registos da amostra obtida com a *Search API* que não foram extraídos pela *Stream API*: 2.5% dos *tweets*, 1% dos utilizadores e 1.3% das *hashtags*. Todavia, a amostra correspondente à *Stream API* é maior, e, consequentemente, contém atividade (utilizadores e *tweets*) que não se encontra presente na mais pequena, obtida com a *Search API*.

Consequentemente, os investigadores retiraram duas conclusões: primeiro, que as redes formadas pelas menções são mais enviesadas para os utilizadores centrais do que as redes formadas por *retweets*, e que este enviesamento poderá estar subestimado se a amostra maior é também ela enviesada para os utilizadores centrais, quando comparado com o fluxo de *tweets* completo do Twitter.

2.3 Conclusão

Dada a tamanha importância e popularidade do Twitter, muito interesse científico foi gerado na sua órbita. De forma expectável, até à data foram realizados inúmeros estudos e projectos com base na sua informação, variando nos objectivos, perspectivas e, consequentemente, nas conclusões. Apesar de distintos, todos partilham uma necessidade comum: a de recolher e reunir dados iniciais, de estabelecer um ponto de partida para uma análise que se espera sólida, e, com isso, atingir a validade das suas conclusões. Como descrito neste capítulo, o Twitter permite o acesso à informação com relativa facilidade, através de métodos da sua API, mas estas recolhas exigem sempre escolhas e decisões por parte dos investigadores. Tudo depende dos recursos, do tempo, mas sobretudo dos objectivos de cada estudo. Qual é o meu objecto de estudo? Toda a Twittosfera? Apenas os utilizadores de língua italiana? Apenas utilizadores residentes nos EUA? Será necessário recolher milhares de *tweets* para analisar comportamentos gerais do Twitter, ou por outro lado apenas pretendo recolher *tweets* que mencionem uma dada figura pública

ou evento? É com base neste tipo de interrogações que é possível determinar que tipos de recolha são mais adequados, já que cada um tem vantagens e desvantagens, mas também diferenças no objectivo que visam. O recurso à API é a opção dominante na extracção dos dados, mas a forma e o método com que esta é utilizada varia.

Na Tabela 2.5 podemos ver os estudos incluídos no estudo bibliográfico, assim como alguns detalhes associados a cada um.

Estudo	*Tweets*	Utilizadores	Tempo de Recolha	Métodos Aplicados	Conjunto base/outro
[JSFT07]	1 348 543	-	1 mês	*Public Timeline*	-
[KGA08]	-	67 527	3 semanas	Seguidos/Seguidores	*Public Timeline*
[HH09]	36 987	-	4 horas (6h,10h,14h,18h)	*Public Timeline*	-
[JZSC09]	150 000	-	13 semanas	*Search API* (Summize)	Palavras
[Cli09]	-	4 414	-	Por ID	-
[RYSG10]	-	200-1000 por tipo	-	Seguidos/Seguidores, *Public Timeline*	hashtags, utilizadores
[SHPC10]	76 milhões	-	39 dias	*Public Timeline* e *Stream API*	-
[CCL10]	29 479 600	1 074 375	5 meses	*Public Timeline*, Seguidos/Seguidores	-
[NPS10]	1 677 978	-	6 meses	Search API	Hashtags
[KLPM10]	106 milhões	41,7 milhões	4 meses	Seguidos/Seguidores e Search API	Perez Hilton,*Trending Topics*
[RDL10]	8 214 019	-	1 semana	Stream API (Spritzer)	-
[WF10]	524 116	-	7 meses e 7 dias	Stream API (Firehose)	-
[BHRG+11]	189 000	-	1 mês	Stream API	Hashtags
[WHMW11]	7 040	-	1 semana	*Public Timeline*	-
[HHSC11]	62 milhões	-	um mês e 10 dias	Stream API (Spritzer)	-
[GCNB+]	5 697 008	-	36 dias	Search API	Palavras
[TGW12]	481 248	1 500	uma semana	*Public Timeline* e Seguidos/Seguidores	-
[CSHS12]	7 milhões	-	um mês	*Public Timeline*	-
[BOM+12]	-	90 000	um ano	Seguidos/Seguidores e Stream API	2 000 utilizadores
[NZBL12]	milhões	-	um ano	Stream (Firehose)	-

Tabela 2.5: Resumo dos vários artigos e trabalhos analisados neste capítulo.

Capítulo 3

Metodologias para Extração de Dados no Twitter

Neste Capítulo serão descritas as metodologias escolhidas para extrair os dados do Twitter. No final do capítulo, são descritos os campos pelo qual um utilizador é definido no presente trabalho.

Como descrito no Capítulo 2, diversas técnicas e metodologias foram elaboradas pela comunidade científica para a extração de dados produzidos no Twitter. Como este estudo pretende comparar algumas dessas técnicas, assim como identificar diferenças (ou semelhanças) na obtenção das amostras de dados, revelou-se como necessária a implementação de um número limitado de técnicas e estratégias. Portanto, e porque é impossível testar todas as técnicas recorrendo a todos os critérios e combinações possíveis, serão explicadas no presente Capítulo todas as decisões tomadas na escolha dos métodos, assim como a implementação de cada um. Concluíu-se que o uso das APIs cedidas pelo Twitter é o recurso mais popular para extrair dados da aplicação. Assim sendo, todas as implementações de métodos de recolha passaram pelo acesso às APIs do Twitter.

Foram assim escolhidos quatro métodos de extração:

- exploração das ligações entre utilizadores ("seguidos" e "seguidores");
- criação de números aleatórios para retirar o utilizador ou tweet com o número de ID correspondente;
- extração dos dados fornecidos através da Stream API.

Verificou-se no Capítulo anterior que estes métodos são os mais populares para a extração do Twitter, com a excepção do recurso ao número de ID do utilizador ou *tweet*, utilizado em apenas um artigo. Os restantes partilham um maior número de presenças em

artigos ou estudos diferentes, com o método de pesquisa baseado nas relações entre utilizadores a verificar-se em 6 estudos; 8 estudos utilizaram a Stream API ou semelhantes. É importante ter em atenção que em muitos casos o detalhe com que se apresenta o método de recolha é demasiado vago, tornando difícil perceber como é que os investigadores extraíram informação do Twitter. Muitos indicam apenas que retiraram os dados da *Public Timeline* através da Twitter API, não especificando o método da API [CSHS12, CCL10]. Tal vem dificultar a identificação correcta da metodologia implementada.

3.1 Considerações Gerais

Como mencionado no ponto anterior, este Capítulo destina-se à descrição dos métodos de recolha de dados implementados no âmbito do presente estudo. Apesar das diferenças próprias de cada estratégia, existe um conjunto de decisões a nível técnico que abrange da mesma forma o desenvolvimento dos métodos. Tais decisões serão justificadas no presente ponto.

No que diz respeito à linguagem de programação a adotar na implementação dos vários métodos, foi crucial optar por uma linguagem flexível e simples, fatores exigidos neste tipo de estudo. Deste modo, uma linguagem de *scripting* revelou-se adequada. Foram testadas três das mais populares linguagens de *scripting* e respectivas bibliotecas de manipulação das funcionalidades das APIs do Twitter. Após um estudo prático das mesmas, e dada a experiência pessoal relativamente limitada em cada uma das três linguagens, concluíu-se que Ruby e Python seriam as linguagens mais apropriadas, dada a sua pequena curva de aprendizagem. Entre as duas, optou-se pela biblioteca *twitter* da autoria de Erik Michaels-Ober [MO] para Ruby. O fato desta se encontrar em constante actualização revelou-se uma mais valia, já que também as funcionalidades das APIs do Twitter são frequentemente alteradas. Esta biblioteca cobre, tanto a REST API, como a Search API do Twitter, mas não possui suporte para os métodos ligados à Stream API. Como tal, decidiu-se usar a biblioteca TweetStream [Int], desenvolvida pela Intridea, Inc, para utilizar as funcionalidades da Stream API.

Independentemente da forma como os dados são recolhidos do Twitter, a gravação dos mesmos de forma persistente é essencial para o estudo e análise da informação retirada. Optou-se por guardar os dados em bases de dados MySQL [Ora], dada a popularidade das mesmas e a experiência pessoal prévia na utilização neste tipo de bases de dados.

3.2 Introdução aos métodos das APIs do Twitter

Na presente secção serão descritos os métodos das APIs do Twitter necessários para a implementação das recolhas descritas neste Capítulo.

API	Método	Argumentos	Resultado	Limite (15minutos)
REST API	*users/lookup* [Twig]	*ids* de utilizador	utilizadores	180
	friends/ids [Twic]	*id* de utilizador, cursor	*ids* de utilizador	15
	followers/ids [Twib]	*id* de utilizador, cursor	*ids* de utilizador	15
Stream API	*statuses/sample* [Twif]	nenhum	*tweet* e utilizador	Não se aplica

Tabela 3.1: Tabela descritiva dos métodos utilizados no presente trabalho.

O método *users/lookup* aceita até 100 números de *id*, devolvendo a informação respectiva a cada um dos utilizadores. Caso todos os *ids* sejam pertencentes a contas de utilizador desconhecidas, suspensas ou eliminadas, não será devolvida qualquer informação relativa a esses *ids*, mas sim um objecto de erro. Caso contrário, será devolvida a informação relativa aos utilizadores válidos.

Os métodos *friends/ids* e *followers/ids* 3.1 devolvem os *ids* correspondentes aos utilizadores presentes na lista de "seguidos" e "seguidores" de um dado utilizador. Este procedimento obriga a alguns cuidados na sua implementação, já que os argumentos dos métodos, assim como as respostas, são limitados. Ambos os métodos devolvem 5 000 números de *id*, sendo necessário efetuar múltiplas invocações de cada método caso um utilizador possua um número de utilizadores nas suas listas de "seguidos" ou "seguidores" superior a 5 000. O cursor para os restantes *ids* da lista é sempre devolvido juntamente com a lista de *ids*, logo é preciso invocar iterativamente o método usando como argumento o cursor devolvido pela chamada anterior.

Os métodos pertencentes à Stream API funcionam de forma díspar dos anteriormente referidos. É criado um *end-point*, onde, iterativamente, o método vai devolvendo *tweets* em tempo real. O *statuses/sample* não necessita de nenhum *input*.

Nem todas as chamadas aos servidores do Twitter devolvem respostas positivas. Na tabela 3.2 podemos observar os erros devolvidos pelas invocações dos métodos da API, detetados durante as recolhas.

Código	Mensagem	Descrição
200	*OK*	Chamada efetuada com sucesso
401	*Unauthorized*	Pedido negado. No caso particular deste trabalho, é devolvido quando se pretende obter informação de um utilizador privado
429	*Too Many Requests*	Foi atingido o limite de chamadas deste método
500	*Internal Server Error*	Problema interno do Twitter
502	*Bad Gateway*	Twitter em baixo ou a ser atualizado
503	*Service Unavailable*	Os serviços do Twitter encontram-se disponíveis, mas com sobrecarga de pedidos
extra	falso *Client Error*	Erro inesperado, atribuido a um problema na compressão da resposta [err]

Tabela 3.2: Tabela das várias respostas aos pedidos HTTP feitos pelos vários métodos de recolha.

3.3 Exploração das ligações entre utilizadores

Este método implica um conjunto de tomadas de decisões determinantes para a forma como a amostra será construída. Como referido no Capítulo 2, este método explora, iterativamente, as listas de "seguidores" e "seguidos" pertencentes ao utilizador. Ambas as listas serão exploradas separadamente, para analisar as características pertencentes aos utilizadores recolhidos por ambos os métodos.

A primeira decisão a ser tomada centra-se na escolha do utilizador ou conjunto base de utilizadores. Verificou-se que esse conjunto deverá ajustar-se ao tipo de estudo que é pretendido. No caso de um estudo que pretenda aglomerar um grande volume de utilizadores indiferenciados do Twitter, normalmente é utilizada uma conta com um número muito elevado de "seguidores"e "seguidos". Por outro lado, quando o estudo pretende reunir utilizadores que partilhem certas caraterísticas específicas, o conjunto base é construído tendo em conta tais características. Neste caso, utilizadores com listas extensas são menosprezados, já que provavelmente serão celebridades ou empresas [RYSG10]. Noutros casos a composição do conjunto base é simplesmente ocultada [CCL10].

No conjunto base para o *script* de amostragem genérica farão parte utilizadores populares no contexto português. Como fatores de escolha, decidiu-se por utilizadores populares que interagem com o Twitter de forma real e direta, que escrevem em português, e que além de um número razoávelmente elevado de "seguidores", também possuam um número relevante de "seguidos". Na Tabela 3.3 podemos ver os utilizadores escolhidos para fazer parte do conjunto base para estes métodos.

username	seguidores	seguidos	tweets
corpodormente	126 657	158	3 384
pedrotochas	18 840	726	3 309
havidaemmarkl	99 867	18 697	8 608
fernandoalvim	74 993	78	3 001
davidfonseca	54 855	166	5 228
brunoaleixo	46 646	34 879	548
pauloquerido	72 500	44 320	98 721

Tabela 3.3: Conjunto base de utilizadores.

A segunda escolha prende-se com o número de utilizadores de cada lista a percorrer, já que lidamos com recursos limitados, tais como o tempo, processamento e as restrições impostas pelo próprio Twitter. Mais uma vez, os investigadores dividem-se quanto a este aspeto. Dos 6 estudos que usaram o método em causa, apenas um efetuou um controlo sobre o limite de utilizadores máximo a pesquisar em cada lista. Deste modo, no presente estudo não será estabelecido qualquer tipo de limite associado ao número de "seguidores" ou "seguidos" de um utilizador a recolher.

A terceira decisão centra-se na inclusão ou exclusão de utilizadores com listas elevadas de "seguidores" ou "seguidos", para a composição da amostra final. Mais uma vez, optou-se pela escolha mais unânime. Dos 6 estudos, apenas 2 excluem utilizadores nesta situação, logo serão dados como válidos utilizadores com qualquer número de "seguidos"ou "seguidores". Como foi referido no Capítulo 2, o método utilizado assenta na combinação de dois métodos da API (*friends/ids* e *followers/ids*) com um terceiro (*users/lookup*).

Para controlar o número de chamadas à API, foi utilizado o campo *followerscount* e *friendscount*, que representam o número de "seguidores" e "seguidos" respectivamente. Este assume o valor de -1 para o primeiro conjunto de ids. Podemos observar com maior detalhe como funciona a recolha de seguidores no Algoritmo 4:

```
Data: integer userid is the numeric identifier of the user being crawled
Data: integer followerscount is the number of followers of the given user
count = followerscount/5000
if count == 0 then
    followers = getFollowers(userid,-1).ids
else
    response = getFollowers(userid,-1)
    followers = response.ids
    for i ← 0 to count do
        response = getFollowers(userid,response.next_cursor)
        followers.concat(response.ids)
    end
    return followers
end
```

Algoritmo 4: Pseudo-código do Algoritmo de extração das listas dos "seguidores" de utilizador.

Depois de obtidas as listas, será necessário extrair a informação relativa aos utilizadores representados pelo *id*. Após a filtragem dos *ids* já presentes na BD, é efetuada a extração da informação relativa aos utilizadores, de 100 em 100 ids.

O método é equivalente para obter os dados relativos aos "seguidos", substituindo o método de recolha de *ids* de "seguidores" pelo método relativo à recolha de *ids* de "seguidos"previamente citado.

Ambas as recolhas decorreram num período de 48 horas. O *script* com base na lista de "seguidores"iniciou a sua execução no dia 26 de Maio de 2013 às 23 horas e 0 minutos. A recolha com base na lista de "seguidos"começou a sua extração no dia 23 às 22 horas e 0 minutos.

3.4 Números de ID aleatórios

Dada a sua simplicidade conceptual, o presente método não apresenta uma grande diversidade de escolhas na sua execução. A primeira fase passa por definir o limite numérico que um ID pode assumir. Tal como DeWitt Clinton procedeu no seu artigo [Cli09], uma maneira simples de determinar o número máximo que um ID pode assumir, passa por registar uma nova conta de Twitter, anotando o ID que a aplicação atribui. Procedeu-se ao registo de uma nova conta de Twitter, para teste. Esta foi registada a 9 de Maio de 2013, às 19 horas, 31 minutos e 11 segundos, obtendo o número de ID 1416190129. De seguida, pretendeu-se recolher a informação, se disponível, correspondente aos números de *id*. Recorremos ao método *users/lookup*.

Sempre que for atingido o limite definido pela API do Twitter, é devolvido um objecto correspondente ao erro. Um dos atributos desse objecto é o tempo de espera necessário até ser possível invocar novamente o método da API, logo o sistema fica em espera durante esse período até que possa continuar a sua execução.

Por fim, foi necessário guardar os dados relativos aos utilizadores válidos devolvidos pelo pedido ao servidor. Caso seja devolvido erro, não se guardam os dados e repete-se o procedimento a partir do segundo passo. Em caso de sucesso, os dados devolvidos pela API são guardados na BD. Neste caso, como pretendemos recolher os dados durante dois dias, todo o processo está contido num ciclo *while*, cuja condição de paragem é o resultado da comparação da data actual com a data de limite pré-establecida. O *script* deu ínicio à sua execução a 9 de Maio de 2013 às 20 horas e 0 minutos, e manteve-se activo durante um período de 2 dias.

No Algoritmo 5 podemos observar com mais detalhe como funciona o método principal do *script* de recolha por ID. Basicamente, o método *lookRandomUsers* retorna um objecto representativo dos vários utilizadores obtidos pelos *ids* gerados aleatóriamente. Estes são gravados na tabela destinada ao registo deste tipo de recolha, cujo o nome é *utilizadorporid*. Caso retorne erro, seja de cliente ou de servidor, esse mesmo erro será registado numa tabela destinada para esse fim.

```
Data: Date Time.now returns the present time
Data: Date @timelimit represents the time limit pre-defined.
while Time.now < @timelimit do
    a = lookRandomUsers
    Database.insertUsers(a,:utilizadorporid)
end
```

Algoritmo 5: Pseudo-código do Algoritmo de recolha de utilizadores pelo campo de ID implementado no presente estudo.

3.5 Dados provenientes da Stream

Como descrito no Capítulo anterior, também os métodos associados à Stream API do Twitter são utilizados com frequência para proceder à recolha de *tweets* e utilizadores. Como tal, para obter a amostra geral de *tweets* fornecidos pela API, é necessário utilizar o método da API *statuses/sample*.

Para definir o *end-point*, utilizamos os recursos da biblioteca mencionada no ínicio do Capítulo para a manipulação da *Stream API*. Este permite definir o conjunto de operação a efetuar por cada iteração do ciclo de recolha. Como pretendemos utilizá-los para análise posterior, necessitamos de guardar persistentemente os dados. Essa operação é precisamente feita a cada iteração do ciclo, inserindo cada utilizador na BD, através do método *insertUserInTable*, que recebe como argumentos o objeto correspondente ao utilizador e o nome da tabela onde os dados deverão ser inseridos.

A recolha de dados iniciou-se dia 23, às 23 horas e 12 minutos, e teve a duração de 24 horas.

No esquema abaixo poderemos analisar com maior detalhe o Algoritmo 6, construído para a extração desta natureza.

```
Data: string table is the name of the table destined to record the stream data
TweetStream::Client.new.sample do |status|
    Database.insertUserInTable(status.user,table)
```

Algoritmo 6: Pseudo-código do Algoritmo para extracção através da Stream API.

3.6 Representação dos dados

Serão analisados apenas os registos correspondentes a utilizadores. Todos os campos para o registo de um utilizador são provenientes dos objetos devolvidos pelo Twitter correspondente à informação do utilizador, exceptuando o campo representativo da data da última publicação de um *tweet*, extraído do objeto referente ao seu último *tweet*. Além dos campos que representam propriedades do utilizador, foram criados outros campos para auxiliarem as análises quanto ao comportamento dos métodos em discussão no presente trabalho. Para este estudo, decidiu-se representar um utilizador através dos seguintes campos descritos na Tabela 3.4.

Campo	Descrição
id de utilizador	Número de identificação do utilizador
Nome da conta	Nome da conta do utilizador
Nome de utilizador	Nome escolhido pelo utilizador
Data de registo	Data de criação da conta do utilizador
Data de entrada	Data de entrada do registo na base de dados
Língua	Língua do perfil do utilizador
Total de "seguidores"	Número total da lista de "seguidores" de um utilizador
Total de "seguidos"	Número total da lista de "seguidos"de um utilizador
Total de listas	Número total de listas do Twitter a que um utilizador pertence
Total de *tweets*	Número total de *tweets* publicados pelo utilizador
Data do último *tweet*	Data de publicação do último *tweet*
Privacidade	Campo que indica se a conta é privada ou pública
Localização	Descrição relativa à localização, introduzida pelo utilizador
URL	hiperligação introduzida pelo utilizador
Fuso horário	Fuso-horário definido automaticamente pelo Twitter
Geolocalização	Campo que indica se a função de geolocalização se encontra ativa ou não
Verificação	Campo que indica se a conta é verificada ou não

Tabela 3.4: Campos do utilizador.

Capítulo 4

Análise de Dados

Neste Capítulo serão descritas as análises aos conjuntos de dados obtidos pelas recolhas definidas no Capítulo anterior. Como explicado previamente, para cada método foram recolhidos registos de utilizadores. Iremos proceder à descrição destas amostras, e para cada campo do registo, será efetuada a análise comparativa dos dados. Para simplificar as referências futuras a cada amostra ou método, serão utilizados os nomes definidos na Tabela 4.1.

Método	Nome da amostra/método
Exploração da lista de "seguidos"	LSeguidos
Exploração da lista de "seguidores"	LSeguidores
Geração de IDs aleatórios	ID
Através da *Sample Stream*	TSample

Tabela 4.1: Referências simplificadas aos métodos implementados e às respetivas amostras.

4.1 Dados gerais das recolhas

Para obter uma percepção geral das características das amostras recolhidas durante o presente estudo, irão ser descritas neste ponto algumas propriedades gerais das amostras de utilizadores. Também o comportamento dos métodos será alvo de análise, através do estudo das invocações dos diferentes métodos das várias API, assim como na quantidade de registos obtidos ao longo do tempo de extração. Na Tabela 4.2 é possível consultar os tamanhos de cada amostra obtida através dos diferentes métodos implementados.

Nome da Amostra	Tamanho
ID	1 393 332
LSeguidos	647 322
LSeguidores	820 863
TSample	2 474 631

Tabela 4.2: Dados gerais das amostras recolhidas.

Para efetuar as comparações estatísticas com base nas amostras de ID, Lista Seguidos e Lista Seguidores, decidiu-se criar uma segunda amostra normalizada, filtrando utilizadores com base na atividade quanto à sua atividade na publicação de *tweets*. Assim sendo, estes métodos terão duas amostras associadas. Na Tabela 4.3 podemos ver os dados relativos a estas amostras normalizadas. Optou-se por cortar todos os utilizadores que publicaram o seu último *tweet* durante todo o mês anterior ao último dia da recolha de dados. Este corte tem como objetivo reduzir as amostras a utilizadores que, tal como os recolhidos através dos restantes métodos, possuam um grau de atividade no Twitter recente, e, assim, excluir todas as contas que possam estar abandonadas ou de uso pouco recorrente.

Nome da Amostra Normalizada	Tamanho da Amostra	% da Amostra Original
IDN	164 320	14,87%
LSeguidosN	489 510	75,62%
LSeguidoresN	372 657	45,40%

Tabela 4.3: Tamanho das amostras normalizadas.

De ressalvar que existiram alguns dados detetados em algumas amostras que se revelaram anormais. Foram detetados quatro registos de utilizador cujas datas de criação de conta correspondiam à data de 01 de Janeiro de 1970. Esta data é conhecida como data Unix, e tanto a data como os restantes campos sugerem que as contas não pertencem a nenhum utilizador de Twitter convencional. Logo, estes registos foram excluídos de qualquer análise. Foi também identificado um registo de utilizador que possuí a data de publicação do último tweet anterior à data de criação da sua conta. Por consequência, este registo foi eliminado da amostra. Por fim, foram detetadas 218 contas na amostra TSample que possuem dados incorretos. Em todos esses registos, o total de *tweets*, o total de "seguidos" e o total de "seguidores" apresentam o valor de -1. Como tal, também estas contas foram removidas da amostra.

4.1.1 Comportamento dos métodos

Nesta secção será analisado o comportamento dos métodos ao longo do tempo de recolha. Serão analisadas tanto a eficiência dos métodos, como as respostas de erro devolvidas pelas chamadas aos métodos da API de cada recolha.

4.1.1.1 Registo dos dados

O que se pretende verificar neste ponto é a frequência com que os dados são guardados na base de dados. Como tal, iremos agrupar as entradas na base de dados verificadas em cada hora de execução do método.

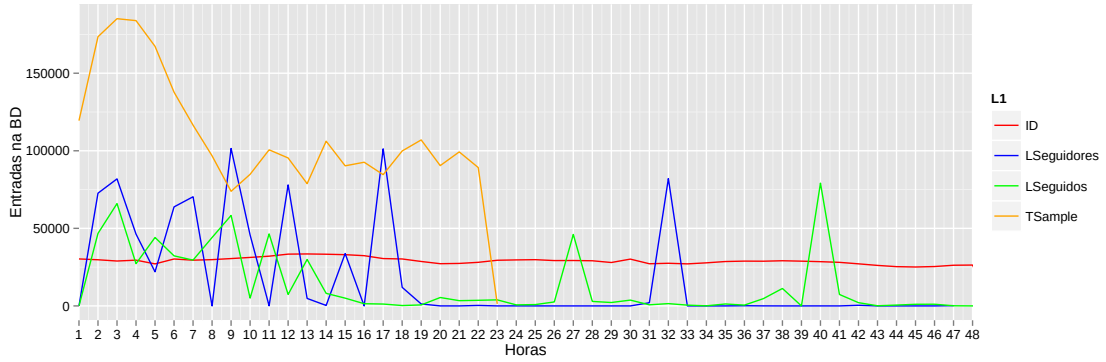

Figura 4.1: Entradas de dados na Base de Dados por hora verificado nos métodos.

Apesar de, em grande maioria, todas as chamadas aos métodos da API serem retornados com sucesso, existem casos em a informação devolvida pelo serviço é falhada, retornando erro. Na Tabela 4.4 podemos verificar as contagens relativas ao tipo de resposta obtida pelas chamadas ao servidor. Dado que os métodos pertencentes à *Stream API* utilizam uma ligação HTTP persistente, foram observados poucos erros.

Método	Chamada	401	429	500	502	503	extra
ID	*users/lookup*	0	0	29	440	3222	3
Lista "seguidores"	*followers/ids*	5	7	0	0	0	0
	users/lookup	0	17	4	7	57	5
Lista "seguidos"	*friends/ids*	23	7	0	0	0	4
	users/lookup	0	16	2	3	27	1
Sample Stream	*stream/sample*	0	0	0	0	0	0

Tabela 4.4: Tabela relativa aos erros retornados por cada invocação do método da API.

Na Tabela 4.5 é possível analisar com maior detalhe as chamadas ao servidor que obtiveram sucesso nas suas respostas. Observa-se que, em média, perto de 54 utilizadores são retornados pelas chamadas efetuadas pelo método ID. Por sua vez, em cada chamada do método *followers/ids*, foram retornados, em média, 3 582 *ids* de utilizador. O método *users/lookup* revela-se mais eficaz nesta estratégia de recolha, devolvendo em média 99,2 utilizadores, muito perto do máximo de 100 utilizadores por chamada. Na estratégia explorativa da lista de "seguidos", as invocações aos métodos da API obtêm menos dados: em média 1 128,25 *ids* devolvidos pelo método *friends/ids*, e 94,6 utilizadores devolvidos pelo método *users/lookup*.

Método	Chamada	Total com sucesso	Média dos dados retornados
ID	*users/lookup*	25 839	53,9236
Lista "seguidores"	*followers/ids*	309	3582,3074
	users/lookup	8 276	99,1861
Lista "seguidos"	*friends/ids*	575	1128,2470
	users/lookup	5 364	94,5899

Tabela 4.5: Tabela relativa às invocações bem sucedidas de cada método da API.

Através da contabilização das entradas de novos registos nas bases de dados, assim como do aproveitamento de cada invocação a métodos da API do Twitter, é possível compreender o funcionamento e eficácia dos métodos empregados, tendo em conta as definições descritas no Capítulo 3. Quanto à obtenção de registos dos utilizadores, todos os métodos revelaram diferenças quanto à quantidade de utilizadores recolhidos por hora, assim como no que diz respeito à quantidade de chamadas da API efetuadas. O método *users/lookup*, quando utilizado na pesquisa aleatória por IDs de utilizador, revelou uma média de aproveitamento proxima de 50%. Tal significa que, em média, por cada 100 *ids* gerados aleatoriamente, apenas metade corresponde a contas ainda presentes no Twitter.

4.2 Utilizadores

Neste ponto irão ser analisados os dados relativos aos utilizadores extraídos pelos métodos implementados no presente trabalho.

4.2.1 Data de Registo de utilizador

Irá proceder-se a verificação da informação relativa ao registo no Twitter das contas recolhidas pelos métodos. Para melhor perceber a evolução da criação de novas contas de Twitter ao longo do tempo, de cada amostra, podemos consultar a figura 4.2.

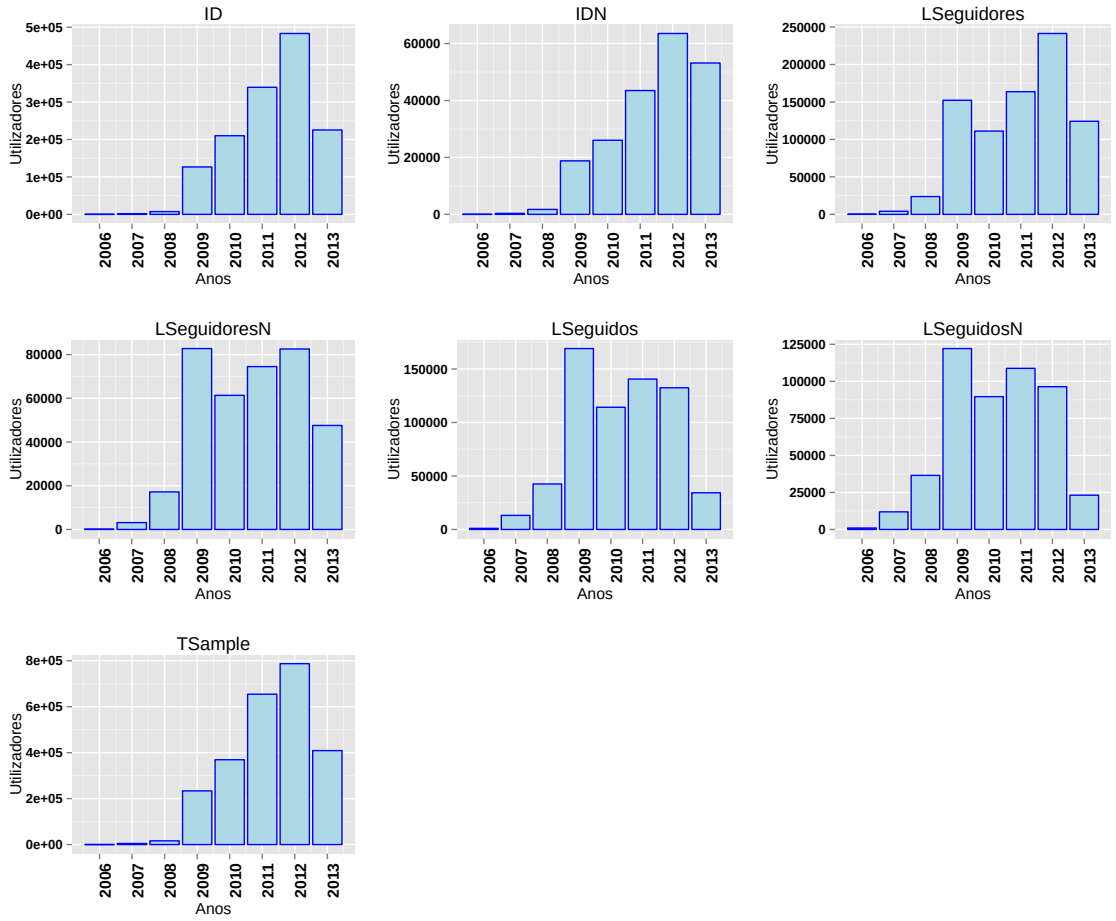

Figura 4.2: Evolução dos registos agrupados por ano.

Apesar das diferenças entre os vários anos de registo da conta de utilizadores variarem de amostra a amostra, praticamente todas as amostras revelam que 2012 foi o ano com maior número de registos novos. A amostra de LSeguidos e a sua subamostra LSeguidosN é a exceção a essa tendência, onde o ano de 2009, seguido do ano de 2011, apresentam mais registos do que o ano de 2012. De ressalvar que o ano presente de 2013 ainda se não se encontra finalizado, sendo natural que este reúna menos contas que o ano anterior.

4.2.2 Tweets

Nesta secção irão ser analisadas as propriedades referentes à publicação de *tweets* por parte dos utilizadores recolhidos nos vários métodos.

O gráfico 4.3 representa a actividade quanto à publicação do último *tweet* por parte dos utilizadores de cada amostra. As subamostras criadas através do corte efetuado com base na data de publicação do último *tweet* possuem todas a mesmo ano de publicação (2013), logo não serão analisadas quanto a este critério, assim como as amostras recolhidas com base nos *tweets*.

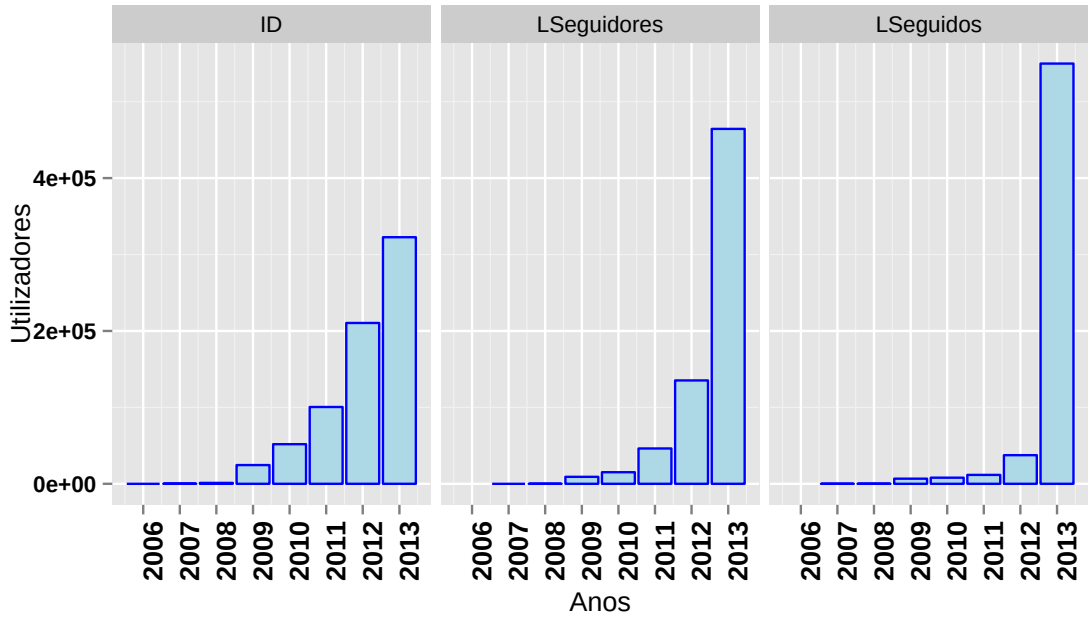

Figura 4.3: Data dos *tweets* emitidos mais recentemente pelos utilizadores das várias amostras, agrupados por ano.

Como se pode facilmente perceber pelo gráfico, grande parte dos utilizadores obtidos pela recolha publicaram o seu último *tweet* no presente ano de 2013, existindo um crescimento ao longo dos anos. No entanto, nota-se uma diferença mais acentuada entre os anos nas recolhas LSeguidores e LSeguidos, sendo que na última a diferença entre 2012 e 2013 é a mais acentuada. Dado que as amostras LSeguidores e LSeguidos percorrem as ligações de amizade de utilizadores, a frequência de amostras com atividade mais recente quanto à publicação de *tweets* é maior face à amostra ID, que assenta numa pesquisa aleatória, e, portanto, capaz de obter ligações com poucas ou nenhumas ligações a utilizadores. A diferença de utilizadores com o último *tweet* publicado em 2012 entre LSeguidores e LSeguidos é considerável. Para as condições definidas para este trabalho, a pesquisa exploratória da lista de "seguidos" mostrou-se mais capaz de obter utilizadores com publicações de *tweets* mais recentes quanto ao seu ano.

4.2.2.1 Número total de tweets

Neste ponto irão ser analisados os totais de *tweets* publicados pelos utilizadores. Os gráficos presentes na Figura 4.4 permitem a análise gráfica da distribuição destes valores nos intervalos de referência definidos.

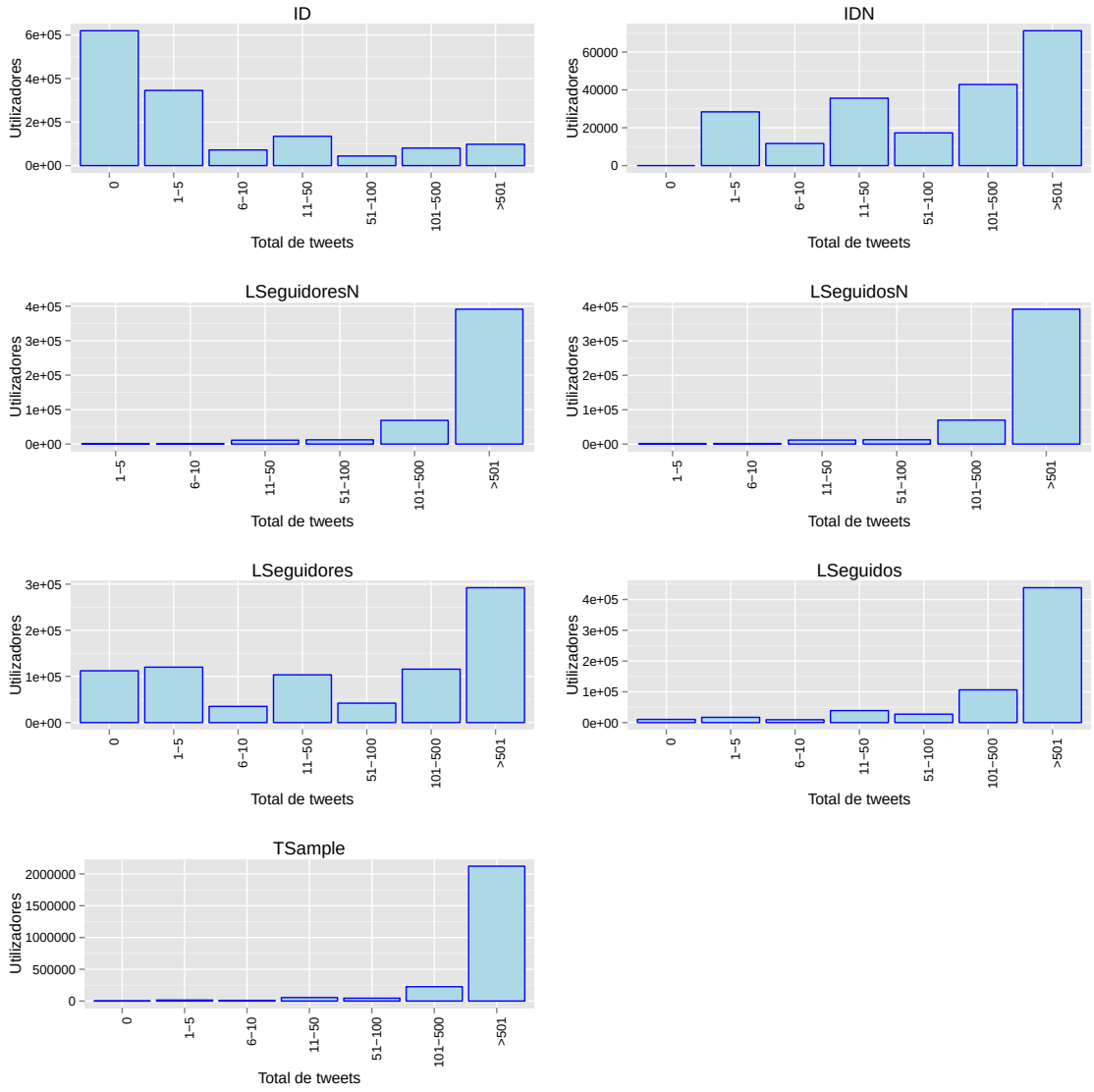

Figura 4.4: Total de tweets emitidos por ano pelos utilizadores recolhidos.

Através da análise da Figura 4.4, pode-se concluir que, de todas as amostras, a TSample reúne a maior parte dos utilizadores com mais do que 500 tweets publicados. A amostra que apresenta um enviesamento semelhante à TSample, excluindo as subamostras, é a de LSeguidos. A ID apresenta um gráfico muito enviesado para valores mais baixos, ao contrário da sua subamostra, que apresenta valores mais distribuídos e com totais de *tweets* mais elevados. As subamostras LSeguidosN e LSeguidoresN apresentam uma distribuição mais semelhante à TSample. As amostras LSeguidores e LSeguidos apresentam diferenças substanciais entre os valores em cada intervalo. A presença de um maior número de registos com menos de 50 *tweets* na amostra LSeguidores, em particular com 0 ou menos de 5 *tweets*, sugere a presença de contas abandonadas ou de uso muito pouco frequente. Este dado vai de encontro aos resultados obtidos na secção anterior, quanto à

data de publicação do último *tweet*.

4.2.2.2 Sumário de valores

Irão ser calculados alguns dados estatísticos referentes a este campo. Na Tabela 4.6 podemos ver os valores respeitantes ao mínimo, máximo, quartis, mediana, média, variância e desvio padrão.

Amostra	Min	1º Q	Mediana	Média	3º Q	Max	Var	DP
ID	0	0	1,0	365,1	11,0	1181716,0	9489386	3080,485
IDN	0	19	149	1937	1057	1181716	53240815	7296,630
LSeguidores	0	4	97	3306	1397	1159952	192518260	13875,095
LSeguidoresN	1	188	1150	6483	5005	1159952	371300354	19269,156
LSeguidos	0	278	1529	6536	5973	2710250	292398601	17099,667
LSeguidosN	1	710	2562	8152	8135	2710250	358305289	18928,954
TSample	0	1264	4672	11703	13564	36440580	1041314587	32269,406

Tabela 4.6: Tabela sumária dos valores totais de *tweet* de cada utilizador.

Pode-se também verificar estes valores gráficamente na Figura 4.5.

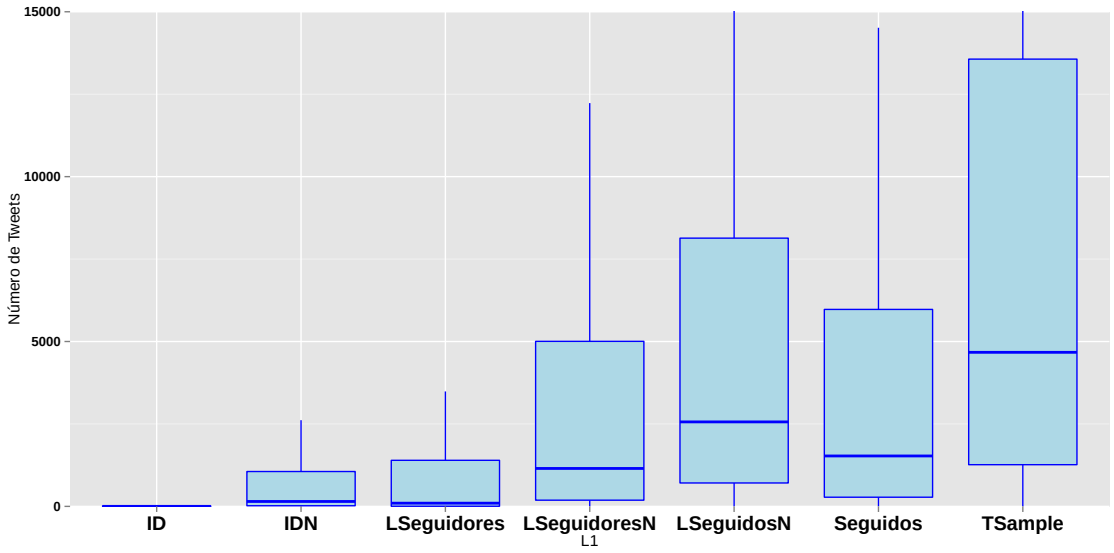

Figura 4.5: Diagrama de caixa e fio relativa aos valores do campo de total dos tweets emitidos por utilizador.

Com base na Tabela 4.6 e no gráfico 4.5, é possível verificar que as amostras apresentam características diferentes entre si. Os valores registados pela amostra ID quanto à centralidade dos dados são muito baixos, verificando uma melhoria significativa na amostra resultante do corte. Esta amostra apresenta-se com características idênticas à de Seguidores, apesar de registar valores diferentes quanto à sua variância. As amostras SeguidoresN e Seguidos também apresentam valores próximos quanto à centralidade dos

valores, apesar de registarem variâncias mais díspares. No entanto, em todas as subamostras construídas com base no corte por atividade, verifica-se um aumento generalizado, tanto nos quartis e na média, assim como na variância.

4.2.3 Média diária de Tweets

Através dos dados presentes nos registos de utilizador obtidos pelas diferentes amostras, podemos calcular a média de *tweets* que cada utilizador publica diariamente. Como tal, recorreu-se à fórmula

$$\frac{tweets\_count}{lastweet\_date - user\_date + 1} \tag{4.1}$$

onde *tweets_count* representa o total de *tweets* de um utilizador, *lastweet_date* a data em dias da publicação do último *tweet* e *user_date* a data em dias da criação da conta do utilizador. É necessário acrescentar mais uma unidade ao denominador, já que a diferença entre ambas as datas deverá ser, no mínimo, um dia, que corresponde a utilizadores que criaram a conta e publicaram o seu último *tweet* no mesmo dia. Como tal, isto é uma média relativa à atividade do utilizador até à publicação do seu último *tweet*.

4.2.3.1 Sumário de valores

Na Tabela 4.7 podemos observar os valores da média de *tweets* de cada utilizador respeitantes ao mínimo, máximo, quartis, mediana, variância e desvio padrão.

Amostra	Min	1º Q	Mediana	Média	3º Q	Max	Var	DP
ID	0	0,0563	0,4091	2,0653	1,0484	978,4820	75,25455	8,675
IDN	0	0,1417	0,6667	3,7150	2,7648	978,4820	141,7276	11,905
LSeguidores	0	0,1538	0,7000	5,1570	3,1402	1039,2 900	362,2017	19,032
LSeguidoresN	0,0006	0,3997	1,6645	7,6976	6,1269	1 039,2900	523,3621	22,877
LSeguidos	0	0,606	2,233	8,404	7,493	11677,000	689,2531	26,254
LSeguidosN	0,0006	0,9874	3,0865	9,7751	9,3353	2 228,8200	526,7359	22,951
TSample	0,000	3,783	11,129	23,048	27,518	26 235,100	1 823,116	42,698

Tabela 4.7: Tabela sumária dos valores das médias diárias de utilizador das várias amostras.

No Figura 4.6 podemos visualizar estes dados gráficamente, através do diagrama de caixa e fio correspondente aos valores expressos na Tabela 4.7.

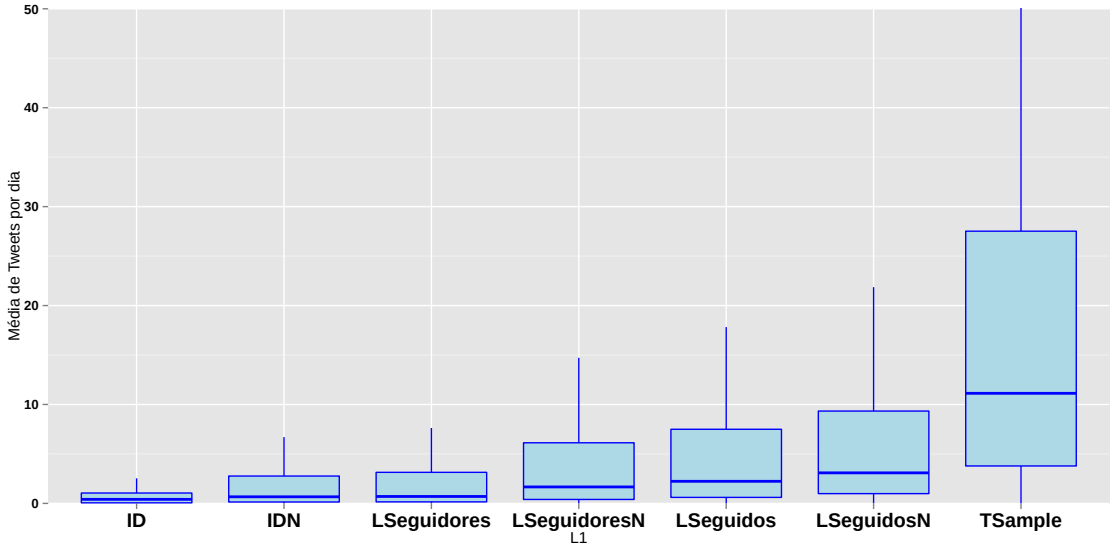

Figura 4.6: Diagrama de caixa e fio relativa aos valores das médias diárias de *tweets* publicados pelos utilizadores.

Através dos dados presentes na Tabela e nos gráficos, pode-se constatar que a amostra TSample destaca-se das outras quanto à centralidade dos seus dados: o seu 1º quartil, de 3.783, é superior ao 3º quartil das amostras de ID, IDN e LSeguidores. Entre as amostras de LSeguidores, LSeguidos e respetivas subamostras, pode-se constatar que, à semelhança dos dados obtidos quanto ao total de *tweets*, também as amostras de LSeguidos apresentam valores ligeiramente superiores em todos os aspectos. O valor correspondente à variância é bastante semelhante nas subamostras.

4.2.4 Seguidores

Neste ponto irá ser estudado o campo relativo aos "seguidores" dos utilizadores recolhidos através dos diferentes métodos. Serão contabilizadas as contagens dos "seguidores", assim como uma análise estatística para cada amostra.

4.2.4.1 Número total de seguidores

Na Figura 4.7 podemos ver os intervalos que reunem os diferentes totais de "seguidores" extraídos de cada amostra.

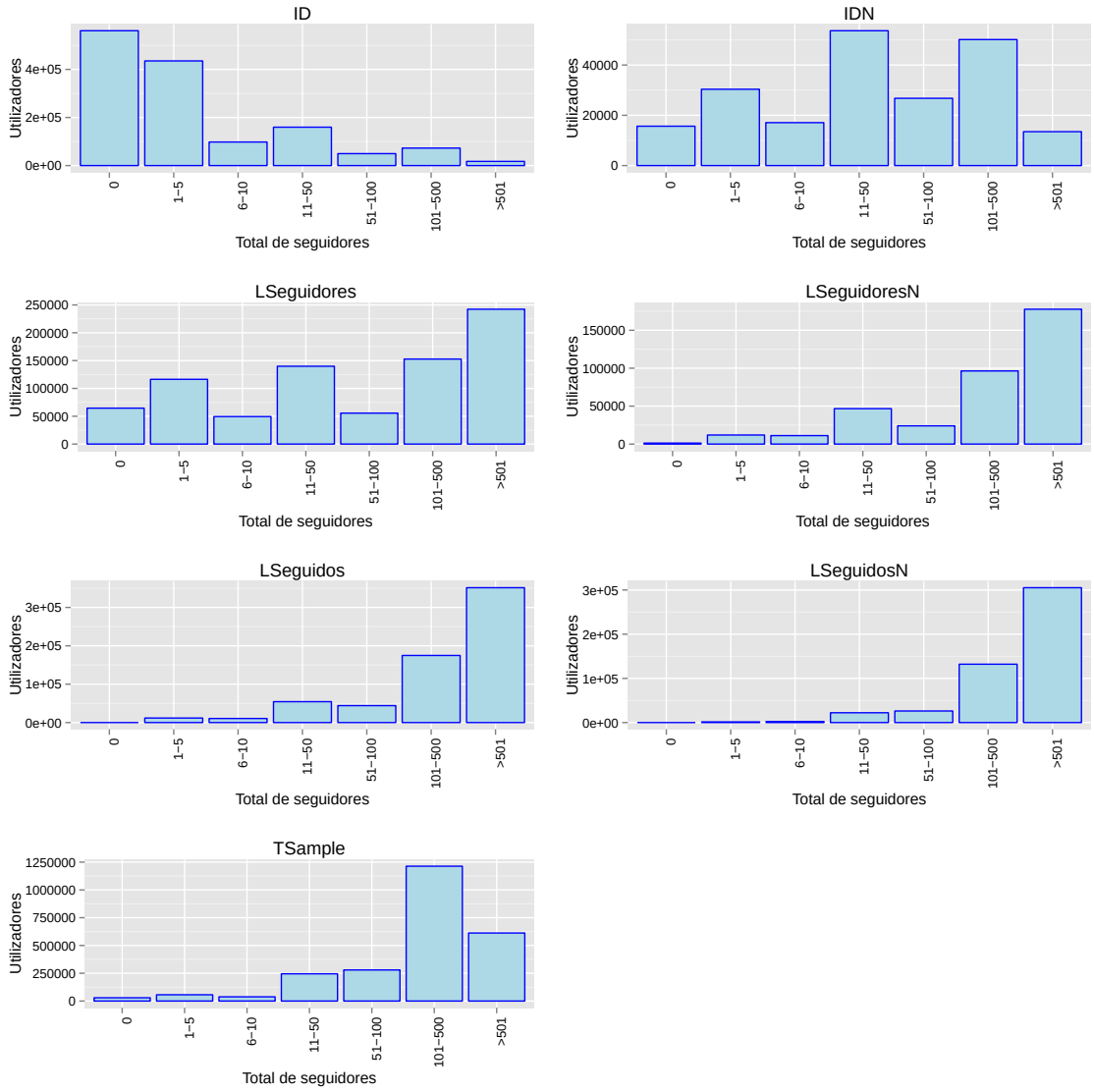

Figura 4.7: Total de seguidores de utilizadores, por definido intervalos.

Na Figura 4.7 pode-se constatar as diferenças que estas amostras apresentam, quanto aos diferentes intervalos de referência definidos. Na amostra TSample, podemos ver que o intervalo entre 100 e 500 seguidores reúne mais utilizadores que os outros intervalos. No entanto, tanto a amostra de LSeguidores, LSeguidoresN, LSeguidos e LSeguidosN apresentam o intervalo de seguidores superior a 500 como o que reúne mais utilizadores. De ressalvar que as subamostras apresentam um enviesamento crescente, revelando que o corte por atividade eliminou da amostra original utilizadores com menos seguidores.

4.2.4.2 Sumário de valores

Presentemente irão ser calculados os valores referentes ao máximo, mínimo, quartis, mediana, média, variância e desvio padrão deste campo.

Amostra	Min	1º Q	Mediana	Média	3º Q	Max	Var	DP
ID	0	0	1,0	58,2	7,0	1 936 850,0	10 518 780	3243,267
IDN	0	7.0	35,0	307,1	134,0	1 936 850,0	69 635 564	8344,793
LSeguidores	0	8	82	2637	702	31 952 594	2 074 145 385	45542,786
LSeguidoresN	0	94	455	5034	1674	31952594	4 564 252 869	67559,255
LSeguidos	0	161	622	17491	3061	39499558	46 560 714 678	215779,319
LSeguidosN	0	267	909	21900	4466	39499558	59 825 296 373	244592,102
TSample	0	95	227	1136	495	7625331	527 241 292	22961,735

Tabela 4.8: Tabela sumária dos valores totais de "seguidores" de cada utilizador.

Trata-se de um campo onde os valores têm diferenças muito significativas entre as amostras. Como observado na análise ao campo de *tweets*, efetuado no ponto anterior, a amostra de ID apresenta valores muito baixos, dificultando a sua análise objetiva. Também subamostra IDN, apesar de um aumento significativo, apresenta valores quanto à sua centralidade e variância mais baixos que as restantes. No entanto, é curioso notar o 1º quartil da amostra ID e LSeguidores praticamente coicide (7 e 8, respetivamente), e as medianas apenas distam 47, o 3º quartil de ambas as amostras apresentam uma diferença de 568. A amostra TSample também apresenta valores de centralidade bastante baixos, contrariando a tendência para valores mais elevados que as amostras de utilizadores com maior atividade revelam, com medianas e valores de 3º quartil na ordem dos milhares. Contudo, tanto a amostra de LSeguidos como a subamostra LSeguidosN destacam-se neste aspeto, apresentando valores bem mais elevados que as restantes quanto aos indicadores de centralidade. Mais uma vez apresentam valores superiores que a amostra LSeguidores. Também na dispersão apresentam os maiores valores.

Na Figura 4.8 é possível visualisar estes valores.

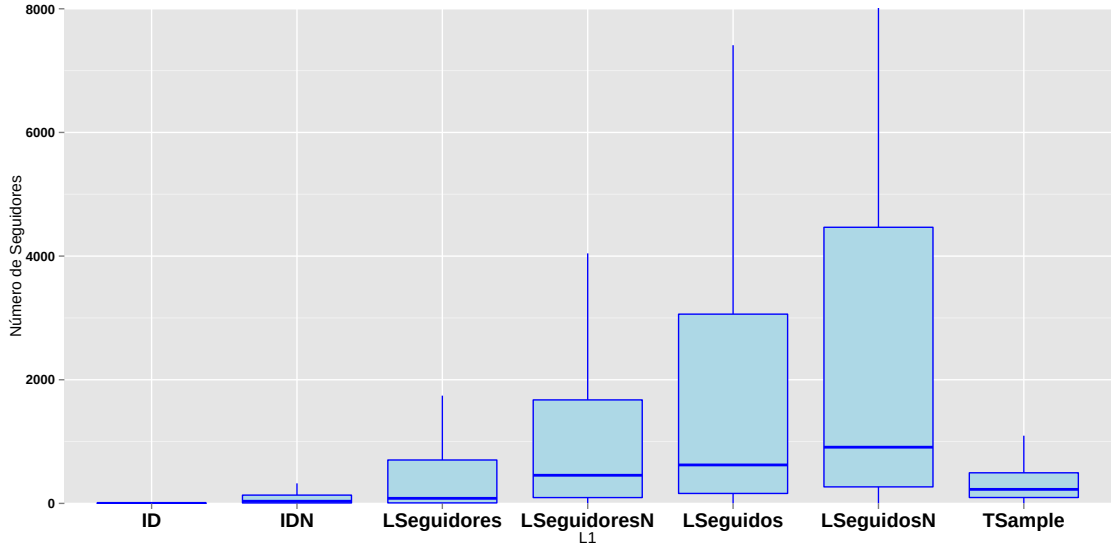

Figura 4.8: Diagrama de caixa e fio relativa aos valores do campo de total dos "seguidores" dos utilizadores.

4.2.5 Seguidos

Neste ponto serão analisados os dados referentes à lista de "seguidos" de cada utilizador, mas específicamente ao totais de "seguidos" recolhidos correspondentes a cada utilizador extraído.

4.2.5.1 Número total de seguidos

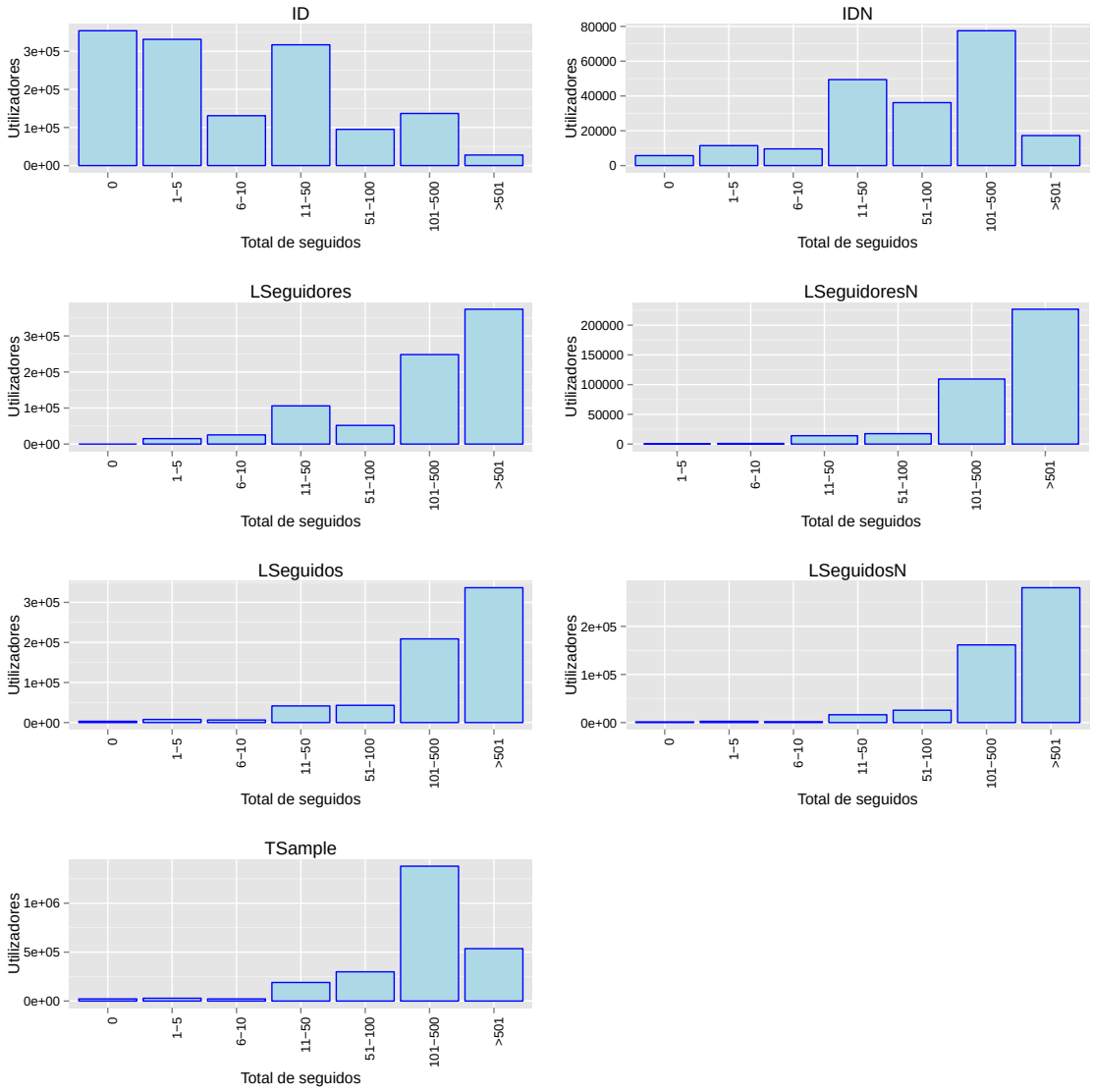

Figura 4.9: Total de seguidos de utilizadores, definido por intervalos.

Através dos valores presentes nos gráficos da Figura 4.9, é possível verificar que as amostras apresentam distribuições diferentes quanto aos totais de "seguidos" de cada utilizador presente na amostra. As amostras LSeguidores e LSeguidos apresentam uma distruibuição semelhante, com excepção da presença de mais utilizadores com número total de "seguidos" entre 11 e 50. No entanto, as suas subamostras revelam-se ainda mais semelhantes quanto à distribuição. A amostra ID apresenta muitos utilizadores com poucos "seguidos", em particular com 0, 1 a 5 e 11 a 50 "seguidos". No entanto, na sua subamostra, estes intervalos possuem frequências de valores mais baixas. O intervalo de 101 a 500 "seguidos" apresenta-se como a moda do gráfico. Quanto à amostra TSample, é possivel verificar que também este intervalo é a sua moda. No entanto, os restantes intervalos possuem, em proporção, menos utilizadores do que os registados na amostra IDN.

4.2.5.2 Sumário de valores

De seguida vamos calcular os valores referentes ao máximo, mínimo, quartis, mediana, média, variância e desvio padrão. Podemos observar estes valores na Tabela 4.9.

Amostra	Min	1º Q	Mediana	Média	3º Q	Max	Var	DP
ID	0	0	6,0	57,5	29,0	372 018,0	267 765,5	517,461
IDN	0	27,0	87,0	213,8	209,0	372 018,0	1 607 049,0	1267,694
LSeguidores	0	110	443	2 310	1 609	1 396 357	143 110 327	11962,873
LSeguidoresN	1	274	830	3 900	1 991	1 396 357	290 783 516	17052,376
LSeguidos	0	184	542	3 031	1 612	1 389 050	203 447 042	14263,486
LSeguidosN	0	252	652	3 533	1 806	1 389 050	252 331 111	15884,934
TSample	0	110,0	222,0	526,2	443,0	986 598,0	10 114 755	3180,370

Tabela 4.9: Tabela sumária dos valores totais de "seguidos" de cada utilizador.

Na Figura 4.10 é possível visualizar estes valores representados através de um diagrama de caixa e fio. Facilmente se conclui que a amostra de ID e IDn contém valores baixos quanto à sua centralidade, fixando o seu 3º quartil ainda antes dos 250. Isto implica que 3 quartos dos utilizadores presentes nestas amostras possuem menos de 250 "seguidos". No entanto, as amostras LSeguidosN e LSeguidoresN apresentam-se bastante semelhantes quanto a este campo. Regista-se uma diferença mais acentuada dos valores quanto à sua centralidade e dispersão entre as amostras de LSeguidores e LSeguidoresN do que as registadas entre as amostras de LSeguidos e LSeguidosN. Neste campo os valores registados nas amostras de LSeguidores e LSeguidos não são tão diferentes entre si do que os registados quanto ao número total de "seguidores".

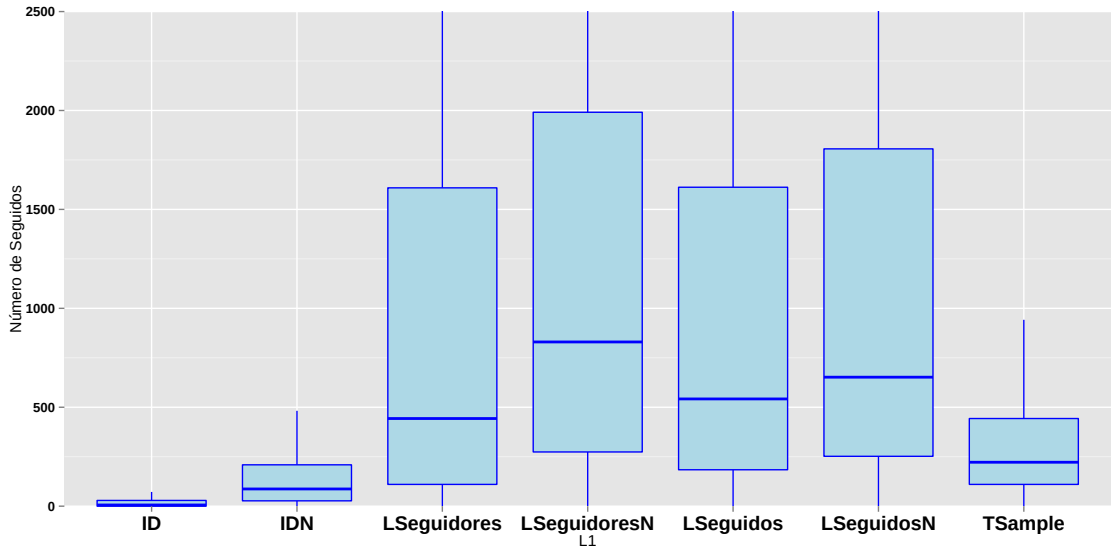

Figura 4.10: Diagrama de caixa e fio relativa aos valores do campo de total dos "seguidos" dos utilizadores.

4.2.6 Língua

Nesta secção irá ser feito o estudo do campo relativo à língua atribuído à conta de um utilizador pelo Twitter.

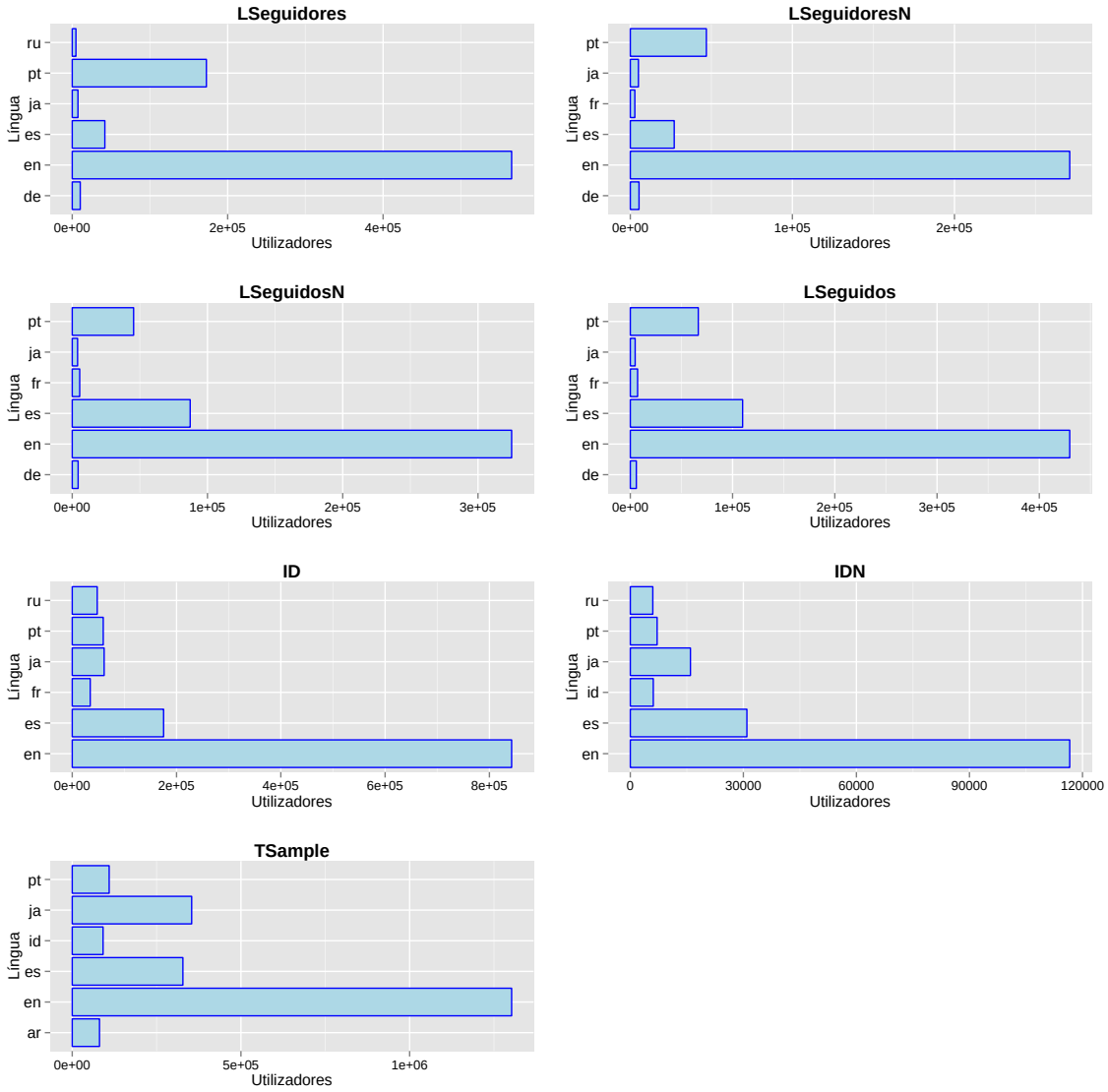

Figura 4.11: Valores do campo *(*lang), correspondente à língua definida para o utilizador.

Como se verifica na Figura 4.11, a língua mais representada em todas as amostras é o inglês ("en"). O espanhol ("es") e o japonês ("ja") surguem em destaque, em particular nas amostras de TSample, ID e IDN. Na amostra de IDN, o indiano ("id") passa a figurar nas 6 línguas mais detetadas, substituindo o francês ("fr") na amostra de ID. Esta língua também surge nas 6 mais representadas registadas na amostra TSample. O russo ("ru") encontra-se muito representado nas amostras de ID, IDN e de LSeguidores, desaparecendo das 6 mais representadas na amostra de LSeguidoresN, onde surge o francês como substituto. O alemão também surge nas línguas com mais utilizadores reunidos,

aparecendo no topo das amostras de LSeguidores e LSeguidos, assim como nas suas subamostras LSeguidoresN e LSeguidosN. A par do inglês e do espanhol, o português ("pt") é a terceira língua que surge nas 6 mais representadas em todas as amostras, se bem que com posições diferentes. Na amostra de LSeguidores e respetiva subamostra, o português é a segunda língua mais representada, seguida do espanhol. No entanto, nas amostras de LSeguidos e LSeguidosN, o português surge atrás do espanhol. Na amostra de ID e IDN, apresenta-se ligeiramente mais representada que o russo, mas menos representada que o japonês. Na amostra TSample, é a terceira língua mais representada, novamente atrás do espanhol.

4.2.7 Privacidade

Quando à privacidade de uma conta de utilizador, serão contabilizadas quantas contas existem definidas como públicas e privadas. Pelos valores da Tabela 4.10, pode-se constatar que existem poucas contas protegidas face ao número de contas públicas. A amostra que apresenta mais contas protegidas é a ID, mas após o corte com base na atividade, não foi registada qualquer conta protegida, sugerindo que todas as contas protegidas pertencem a contas com graus de atividade muito baixos, ou possívelmente contas abandonadas. Também nas amostras de LSeguidoresN e LSeguidosN se nota uma drástica redução para perto de 0.0% quanto às contas protegidas, face às milhares de contas com esta propriedade ativa registadas nas amostras originais.

Amostra	Protegidas		Públicas	
	Total	%	Total	%
ID	74 828	5,37%	1 318 504	94,63%
IDN	0	0,0%	207 123	100,0%
LSeguidos	23 021	3,56%	624 301	96,44%
LSeguidosN	3	0,00%	489 507	100,00%
LSeguidores	40 140	4,89%	780 723	95,11%
LSeguidoresN	3	0,0%	369 174	100,0%
TSample	0	0,0%	2 474 849	100,0%

Tabela 4.10: Contagem das contas públicas e privadas de utilizador.

4.2.8 Localização

O campo de localização tem sido alvo de diversos estudos. Apenas será verificado quantos utilizadores decidem preencher este campo, e quantos utilizadores optam por deixá-lo vazio. Foram contabilizadas todas as contas com valores nulos ou com a string ” referentes a este campo.

As amostras apresentam valores diferentes entre si quanto a este campo. Na amostra de ID, maioria das contas apresentam o campo de localização por preencher, mas após

Amostra	Com localização		Sem localização	
	Total	%	Total	%
ID	331 123	23,76%	1 062 209	76,24%
IDN	106 326	51,33%	100 797	48,67%
LSeguidos	488 335	75,44%	158 987	24,56%
LSeguidosN	387 648	79,19%	101 862	20,81%
LSeguidores	539 198	65,69%	281 665	34,31%
LSeguidoresN	288 054	78,03%	81 123	21,97%
TSample	1 502 154	60,70%	972 477	39,30%

Tabela 4.11: Contagem das contas com localização definida no perfil de utilizador.

o corte por atividade, o número de contas com o campo de localização preenchido e o número de contas com o campo vazio é praticamente o mesmo. Nas amostras de LSeguidos e LSeguidores, apesar de maioria dos registos de utilizador se encontrarem com este campo preenchido, o número de contas registadas com o campo vazio é ainda assim bastante relevante. Na amostra TSample, cerca de 60% da amostra corresponde a utilizadores com este campo preenchido. Estes valores poderão ser consultados com maior detalhe na Tabela 4.11.

4.2.9 URL

Irá ser apurado neste ponto quantos utilizadores introduzem dados no campo destinado a conter um *URL*.

Amostra	Com URL		Sem URL	
	Total	%	Total	%
ID	99 764	7,16%	1 293 568	92,84%
IDN	40 531	20,00%	166 592	80,00%
LSeguidos	365 154	56,41%	282 168	43,59%
LSeguidosN	301 375	61,57%	188 135	38,43%
LSeguidores	310 710	37,85%	510 153	62,15%
LSeguidoresN	197 352	53,45%	171 825	46,54%
TSample	666 011	26,91%	1 808 838	73,08%

Tabela 4.12: Contagem das contas com *URL* definido no perfil de utilizador.

Verificam-se algumas diferenças entre as amostras quanto ao preenchimento deste campo. Na amostra por ID, verificamos que a esmagadora maioria dos utilizadores não introduziram qualquer informação neste campo. No entanto, após o corte, a diferença entre os dois valores apresenta-se mais equilibrado, revelando que maioria das contas que apresentam este campo vazio pertencem a utilizadores que não publicam *tweets* recentemente. Na Tabela 4.12 podemos verificar todas as contagens efetuadas quanto ao preenchimento deste campo.

4.2.10 Fuso horário

Outro dos campos associados a uma conta de Twitter é o seu fuso horário, determinado manualmente pelo utilizador no seu perfil. Irão ser estudados apenas os 6 fuso horários que reuniram mais utilizadores nas amostra obtidas.

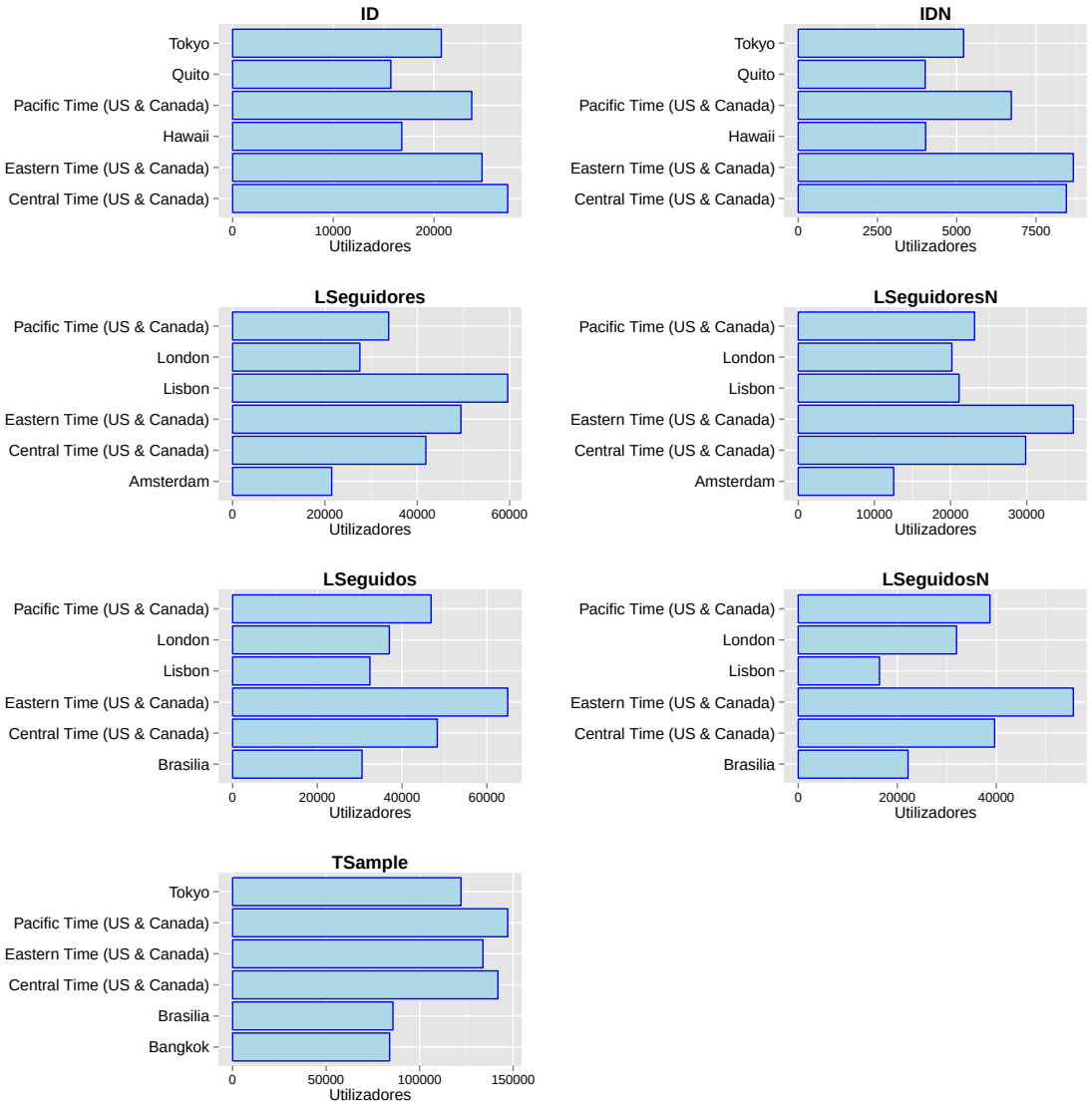

Figura 4.12: 6 fuso horários mais representados nas amostras de utilizadores.

Verifica-se que existem 3 fusos horários presentes em todas as amostras, correspondentes aos Estados Unidos da America e Canadá. Além da presença em todas as amostras, são os fusos horários que reúnem maior número de utilizadores. Nas amostras de LSeguidores e LSeguidos encontram-se também muitos utilizadores com o fuso horário correspondente a Lisboa e Londres. No entanto, na amostra de LSeguidores o 6 lugar é ocupado por Amesterdão, enquanto na amostra de LSeguidos esse mesmo lugar é ocupado por Brasilia. De notar que as subamostras resultantes destas amostras mantêm os

mesmos 6 fusos horários, mas reduzem significamente a proporção de utilizadores com fuso horário atribuído a Lisboa. O fuso horário de Brasilia também se encontra nos 6 mais registados pela amostra de TSample, com praticamente o mesmo número de utilizadores com fuso horário atribuído a Banguecoque. Também o fuso horário correspondente a Tóquio se encontra nos fusos horários mais detetados por esta amostra, ocupando também um lugar cimeiro na amostra de ID, e consequentemente, na subamostra IDN. Quito e Havai são os outros dois fusos horários mais registados nas amostras de ID e IDN. Na Tabela 4.12 encontram-se toda a informação relativa aos dados deste este campo.

4.2.11 Geolocalização

Para a análise quando a este campo, irão ser contabilizados os utilizadores que se encontram com a opção de geolocalização ativada ou desativada.

Amostra	Com Geolocalização		Sem Geolocalização	
	Total	%	Total	%
ID	119 060	8,55%	1 274 272	91,45%
IDN	47 172	22,78%	159 951	77,22%
LSeguidos	202 692	31,31%	444 630	68,69%
LSeguidosN	173 155	35,37%	316 355	64,63%
LSeguidores	158 801	19,35%	662 062	80,65%
LSeguidoresN	108 540	29,40%	260 637	70,60%
TSample	925 677	37,40%	1 549 172	62,60%

Tabela 4.13: Contas de utilizador com geolocalização ativa ou inativa.

Tanto na amostra de ID como na sua subamostra IDN, é possível verificar que existem mais utilizadores com geolocalização desativada, apesar da diferença proporcional diminuir considerávelmente da amostra original para a sua subamostra. Esta tendência verifica-se nas restantes amostras de LSeguidos e LSeguidores, onde se verifica um maior decréscimo de utilizadores com geolocalização inativa face aos que possuem esta opção ativada. Também na amostra TSample é possível verificar que a maioria dos utilizadores não optam por ativar esta propriedade. Na Tabela 4.13 é possível verificar estes dados com maior grau de promenor.

4.2.12 Verificação

Neste ponto as amostras serão analisadas quanto à verificação das contas obtidas por cada amostra.

Amostra	Contas verificadas		Contas não verificadas	
	Total	%	Total	%
ID	124	0,01%	1 393 208	99,99%
IDN	116	0,06%	207 007	99,94%
LSeguidos	19 089	2,95%	628 233	97,05%
LSeguidosN	18 384	3,76%	471 126	96,24%
LSeguidores	1 915	0,23%	818 948	99,77%
LSeguidoresN	1 857	0,50%	367 320	99,50%
TSample	2 341	0,09%	2 472 508	99,91%

Tabela 4.14: Contas de utilizador com verificação e sem verificação.

Como se pode observar pelos dados da Tabela 4.14, todas as amostras apresentam números de contas não verificadas mais elevados face às contas verificadas. Em todas as amostras as contas verificadas surgem com valores resíduais, tendo em conta os tamanhos das amostras. No entanto, é interessante verificar que nas amostras LSeguidos e LSeguidosN estes valores são largamente superiores ao das outras amostras. Outro aspeto interessante prende-se com a pouca diferença registada entre as amostras e as subamostras quanto aos utilizadores com contas verificadas, sugerindo que estes utilizadores serão, por norma, muito ativos quanto à publicação de *tweets*.

4.2.13 Cobertura de dados

Neste ponto irão ser descritas as intersecções de dados obtidos entre algumas das amostras.

4.2.13.1 LSeguidores e LSeguidos

As amostras de LSeguidores e LSeguidos são construídas com base no mesmo conjunto base de utilizadores. Como tal, irá ser verificada a sobreposição de alguns dos seus dados.

Amostra	Total de utilizadores	% de utilizadores
LSeguidores	820 863	14,75%
LSeguidos	647 322	18,71%
Cobertura	121 098	-

Tabela 4.15: Número de utilizadores presentes em ambas as amostras.

Na Tabela 4.15 verifica-se que, incluindo os 7 utilizadores que consistiram na base de utilizadores para a recolha, 121 098 utilizadores foram recolhidos por ambos os métodos.

Contabilizaram-se as relações recíprocas detetadas nas amostras, ou seja, relações em que os utilizadores são simultâneamente "seguidores" e "seguidos" de outro utilizador.

Relação	Total de relações
Seguidores	1 106 491
Seguidos	814 195
Recíprocas	70 499

Tabela 4.16: Relações entre os utilizadores das amostras LSeguidos e LSeguidores.

Foram detetadas 70 499 relações desta natureza. Podemos consultar estes valores na Tabela 4.16.

4.2.14 Intervalos de Confiança

Nesta secção iremos calcular os intervalos de confiança de cada amostra, correspondente aos campos relativos ao número total de *tweets*, total de "seguidores" e total de "seguidos". Como grau de confiança, irá ser utilizado o valor de 95%.

Para que o cálculo dos intervalos de confiança seja robusto, é necessário que se verifiquem as seguintes condições:

- número de observações independentes;
- número de observações suficientemente grande para que a média da amostra tenda a uma distribuição Normal;

Ambas as propriedades se verificam, já que as observações em cada amostra são independentes e com ordens de grandeza na ordem dos milhares e milhões. Como é possível

Amostra	IC 95%		
	Número de *tweets*	Número de "seguidores"	Número de "seguidos"
ID	616,2238 – 635,3471	97,46596 – 118,54359	95,67741 – 99,01584
LSeguidores	3 894,473 – 3 967,153	3 077,007 – 3 318,126	2722,692 – 2785,757
LSeguidos	6 723,502 – 6 810,76	17 837,54 – 18 945,96	3124,885 – 3198,022
TSample	11 662,53 – 11 742,94	1 107,292 – 1 164,512	522,2888 – 530,2142

Tabela 4.17: Valores dos intervalos de confiança calculados relativos a diferentes campos de cada amostra.

verificar na Tabela 4.17, os intervalos de confiança para cada amostra apresentam valores muito diferentes para cada campo, entre as várias amostras. Estes intervalos nunca se intersectam, sendo possível concluir a sua diferença com um grau de confiança de 95%.

Capítulo 5

Conclusões e Trabalho Futuro

Após a análise das amostras obtidas no decorrer do presente trabalho, serão descritas neste Capítulo as conclusões destas retiradas. Serão igualmente apresentados ideias e percursos a adotar em trabalhos futuros no âmbito da análise e caracterização de amostras do Twitter.

5.1 Conclusões

Este trabalho teve como premissa inicial responder à pergunta: técnicas de recolha de dados diferentes originam amostras diferentes? Nesta secção serão descritas as diferenças mais relevantes identificadas nas amostras, assim como as semelhanças, que confirmam algumas propriedades gerais sobre a rede social.

Através do cálculo dos intervalos de confiança referentes às médias do total de *tweets*, "seguidores" e "seguidos", é possível concluir que todos os métodos determinaram intervalos para as médias reais da população muito díspares. Os intervalos são tão díspares que nunca se intersetam. Ou seja, com um grau de confiança de 95%, nunca iremos obter amostras cuja as médias sejam identicas para os três campos analisados. Este dado evidência a grande diferença que estas amostras revelam quanto a estes campos.

No que diz respeito ao comportamento dos métodos utilizados, observaram-se características diferentes entre os mesmos. O método de ID representa pouca variação na sua eficiência, devolvendo utilizadores correspondentes a cerca de 50% dos *ids* por cada chamada ao método da API; as recolhas de utilizadores através das listas de "seguidores" e "seguidos" diferem na eficiência dos seus métodos. As invocações efetuadas aos métodos da API presentes no primeiro devolvem, em média, mais dados, ao contrário do segundo, em que a média dos dados devolvidos é inferior.

Todas as amostras presentes no trabalho revelaram uma tendência para a existência de um maior número de utilizadores com mais "seguidos" do que "seguidores". Tal vem influênciar a exploração por essas listas de utilizadores, já que com menos utilizadores presentes na lista de "seguidores", são devolvidos menos números de *ids* por chamada. Também o número médio de utilizadores devolvidos pelo método da API *users/lookup* se apresenta menor. Por outro lado, a exploração da lista de "seguidores" revela um maior aproveitamento, devolvendo mais *ids* de utilizador, assim como mais utilizadores com o método *users/lookup*.

O método Sample da *Stream API* revela, quanto à obtenção de registos, um comportamento variável ao longo do tempo, já que depende do fluxo da atividade registada quanto à publicação de *tweets*, também variável ao longo do tempo [MPLC13].

Foi possível observar uma concordância entre as amostras em aspetos como a língua mais utilizada no perfil dos utilizadores. O inglês é predominante, seguido do japonês e do espanhol, muito menos relevantes mas igualmente presentes em todas as amostras. Este aspeto ganha ainda mais relevância na verificação dos dados do fuso horário, onde surgem nos lugares cimeiros Tokyo e três fusos horários americanos. Lisboa surge em destaque nas amostras LSeguidores e LSeguidos, já que o conjunto base da pesquisa exploratória partilha este fuso horário; todas as amostras apresentam uma predominância de utilizadores que se encontram com a opção de geolocalização desativada; ao mesmo tempo, a percentagem de contas não verificadas são predominantes em todas as amostras. No entanto, a amostra de LSeguidos apresentou números de contas verificadas superiores às restantes. Enquanto nas restantes amostras se registam entre 116 a 2 341, que correspondem a nunca mais de 0,50% da totalidade dos utilizadores, na amostras de LSeguidos foram recolhidos 19 089 utilizadores com contas verificadas, que correspondem a cerca de 3% da amostra.

Os métodos baseados na exploração das ligações de amizade dos utilizadores revelaram diferenças entre si. A Amostra de LSeguidores apresentou valores substancialmente mais altos nos campos: total de *tweets*, média de *tweets* e total de "seguidores". A combinação destes fatores sugere a presença de utilizadores com maior atividade no Twitter na amostra LSeguidos, assim como utilizadores cujos *tweets* despertam maior interesse noutros utilizadores. Uma explicação possível centra-se na maior probabilidade de existirem mais contas abandonadas ou falsas presentes nas listas de "seguidores", do que nas listas de "seguidos". Estas contas irão apresentar números reduzidos (ou nulos) de "seguidores".

Muitas vezes adotados em conjunto, o estudo da sobreposição entre os utilizadores devolvidos por ambos reenforça a utilidade dessa opção, já que se verificou que a apenas 14,75% e 18,71% de cada amostra representam os mesmos utilizadores. O recurso a ambos os métodos, irá, presumivelmente, complementar as amostras com utilizadores com características que, de outra forma, não seriam recolhidos.

A amostra através da geração de IDs aleatórios de utilizador apresenta propriedades

que mais nenhum outro método de recolha estudado apresenta, sob as premissas e condicionantes definidas no Capítulo 3. Apesar desta diferença, reflete as propriedades mais gerais observadas também nas restantes amostras. No entanto, apenas 14,87% da amostra corresponde a utilizadores com um grau de atividade quanto à publicação de *tweets* aceitável. A simplicidade da sua implementação e a forma de chegar a utilizadores com poucas conexões faz com que seja uma abordagem interessante e diferente, raramente adotada pela comunidade científica.

5.2 Trabalho Futuro

Como trabalho futuro, não só amostras compostas por utilizadores, mas também as amostras de *tweets* podem ser alvo de estudo quanto a possíveis diferenças que possam apresentar, derivadas de diferentes metodologias e recolhas. Quanto à diferença na qualidade de amostras de utilizadores, seria do interesse científico implementar as mesmas recolhas exploratívas sob condições diferentes (conjuntos base com características diferentes e mais tempo de recolha), para possibilitar a descoberta, não só das diferenças entre as amostras, mas também para decifrar o impacto que essas decisões implicam nos trabalhos finais.

Referências

[Ale] Alexa. Top sites by category: Computers/internet/on the web/online communities/social networking. http://www.alexa.com/topsites/category/Computers/Internet/On_the_Web/Online_Communities/Social_Networking. Acedido em: 19/07/2013.

[BHRG+11] Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iñiguez, María Pilar Pérez, Gonzalo Ruiz et al. Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study. *PLoS One*, 6(8):e23883, 2011.

[BOM+12] M. Boanjak, E. Oliveira, J. Martins, E. Mendes Rodrigues e L. Sarmento. Twitterecho: a distributed focused crawler to support open research with twitter data. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 1233–1240. ACM, 2012.

[CCL10] Zhiyuan Cheng, James Caverlee e Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 759–768, New York, NY, USA, 2010. ACM.

[Cli09] DeWitt Clinton. Sampling Twitter. Janeiro 2009.

[Cra13] Kate Crawford. The hidden biases in big data. http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html, 2013.

[CSHS12] J. Cranshaw, R. Schwartz, J.I. Hong e N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. *Association for the Advancement of Artificial Intelligence*, 2012.

[err] getting -sometimes- twitter::Error::Clienterror: end of file reached. https://github.com/sferik/twitter/issues/370. Acedido em: 27/05/2013.

[GBWR+12] S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer e Y. Moreno. Assessing the bias in communication networks sampled from twitter. *Available at SSRN 2185134*, 2012.

[GCNB+] Christophe Giraud-Carrier, E Shannon Neeley, Michael Dean Barnes, Kyle Prier, Joshua Heber West, Parley Cougar Hall e Carl Lee Hanson. Temporal variability of problem drinking on twitter. *Journal of Biophysical Chemistry*, 2.

[GNIa] GNIP. http://www.gnip.com/. Acedido em: 29/10/2012.

[GNIb] GNIP. Twitter Spritzer feed. http://www.gnip.com/twitter/spritzer. Acedido em: 29/10/2012.

[HH09] C. Honey e S.C. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pages 1–10. IEEE, 2009.

[HHSC11] B. Hecht, L. Hong, B. Suh e E.H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 237–246. ACM, 2011.

[Int] Inc. Intridea. intridea / tweetstream. https://github.com/intridea/tweetstream. Acedido em: 27/01/2013.

[JSFT07] Akshay Java, Xiaodan Song, Tim Finin e Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[JZSC09] Bernard J Jansen, Mimi Zhang, Kate Sobel e Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[KGA08] Balachander Krishnamurthy, Phillipa Gill e Martin Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, WOSP '08, pages 19–24, New York, NY, USA, 2008. ACM.

[KKNG12] J. Kulshrestha, F. Kooti, A. Nikravesh e K.P. Gummadi. Geographic dissection of the twitter network. In *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[KLPM10] H. Kwak, C. Lee, H. Park e S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[MO] Erik Michaels-Ober. sferik / twitter. https://github.com/sferik/twitter. Acedido em: 27/01/2013.

[MPLC13] Fred Morstatter, Jürgen Pfeffer, Huan Liu e Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *The 7th International Conference on Weblogs and Social Media (ICWSM-13), Boston, MA.*, 2013.

[Nat] Daniel Nations. The top social networking sites. http://webtrends.about.com/od/socialnetworking/a/social_network.htm. Acedido em: 29/10/2012.

[NPS10] M. Nagarajan, H. Purohit e A. Sheth. A qualitative examination of topical tweet and retweet practices. In *International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[NZBL12] M. Naaman, A.X. Zhang, S. Brody e G. Lotan. On the study of diurnal urban routines on twitter. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[Ora] Oracle. MySQL. http://www.mysql.com/. Acedido em: 27/01/2013.

[RDL10] Daniel Ramage, Susan Dumais e Dan Liebling. Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*, volume 5, pages 130–137, 2010.

[RYSG10] Delip Rao, David Yarowsky, Abhishek Shreevats e Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[Sem] Semiocast. Twitter reaches half a billion accounts more than 140 millions in the u.s. http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US. Acedido em: 29/10/2012.

[SHPC10] Bongwon Suh, Lichan Hong, P. Pirolli e E.H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 177 –184, August 2010.

[spi] Spinn3r. http://www.spinn3r.com/. Acedido em: 29/10/2012.

[TBP11] M. Thelwall, K. Buckley e G. Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

[TGW12] Y. Takhteyev, A. Gruzd e B. Wellman. Geography of twitter networks. *Social Networks*, 34(1):73–81, 2012.

[twia] Twitter Portugal. http://twitterportugal.com/. Acedido em: 29/10/2012.

[Twib] Twitter. GET followers/ids. https://dev.twitter.com/docs/api/1.1/get/followers/ids. Acedido em: 01/12/2012.

[Twic] Twitter. GET friends/ids. https://dev.twitter.com/docs/api/1.1/get/friends/ids. Acedido em: 01/12/2012.

[Twid] Twitter. GET search/tweets. https://dev.twitter.com/docs/api/1.1/get/search/tweets. Acedido em: 29/10/2012.

[Twie] Twitter. GET statuses/publictime_line. https://dev.twitter.com/docs/api/1/get/statuses/public_timeline. Acedido em: 19/02/2013.

[Twif] Twitter. GET statuses/sample. https://dev.twitter.com/docs/api/1.1/get/statuses/sample. Acedido em: 19/02/2013.

[Twig] Twitter. GET users/lookup. https://dev.twitter.com/docs/api/1.1/get/users/lookup. Acedido em: 01/12/2012.

[Twih] Twitter. GET users/show. https://dev.twitter.com/docs/api/1.1/get/users/show. Acedido em: 01/12/2012.

[Twii] Twitter. Search API. https://dev.twitter.com/docs/api/1/get/search. Acedido em: 29/10/2012.

[Twij] Twitter. The Streaming APIs. https://dev.twitter.com/docs/streaming-apis. Acedido em: 29/10/2012.

[Twik] Twitter. Twitter Decahose feed. http://www.gnip.com/twitter/decahose. Acedido em: 29/10/2012.

[Twil] Twitter. Twitter Entities. https://dev.twitter.com/docs/tweet-entities. Acedido em: 29/10/2012.

[Twim] Twitter. Twitter Firehose. https://dev.twitter.com/docs/api/1/get/statuses/firehose. Acedido em: 29/10/2012.

[Twin] Twitter. Twitter rate limiting. https://dev.twitter.com/docs/rate-limiting. Acedido em: 29/10/2012.

[Twio] Twitter. The twitter REST API. https://dev.twitter.com/docs/api. Acedido em: 29/10/2012.

[Twip] Twitter. User Streams. https://dev.twitter.com/docs/streaming-apis/streams/user. Acedido em: 29/10/2012.

[Twiq] Twitter. Using the Twitter Search API. https://dev.twitter.com/docs/using-search. Acedido em: 24/03/2012.

[WF10] Stina Westman e Luanne Freund. Information interaction in 140 characters or less: genres on twitter. In *Proceedings of the third symposium on Information interaction in context*, pages 323–328. ACM, 2010.

[WHMW11] S. Wu, J.M. Hofman, W.A. Mason e D.J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714. ACM, 2011.